
# A Vision-Based Approach For Assisting Functional Assessment Involving Activities of Daily Living

A thesis submitted to the Edge Hill University for the degree of
Doctor of Philosophy
in the Faculty of Arts and Sciences

May 2021

Bappaditya Debnath
Department of Computer Science

# Acknowledgements

I want to express my sincere gratitude to my supervisory team Dr Ardhendu Behera (Director of Studies), Dr Swagat Kumar and Prof Mary O'Brien. I would like to especially thank Dr Ardhendu for his continuous and very involved support of the research. The research output that I was able to produce would not have been possible without his continuous involvement, encouragement, motivation and inspiration. I also thank Dr. Motonori Yamaguchi for who was part of my supervisory team during the first two years of the research. Special thanks to Dr Helen Carey, who provided valuable guidance for the creating the dataset. I am grateful to the volunteers for taking part in the dataset and Pradeep for painfully proof-reading the document. Finally, I would also thank Prof Nik Bessis, Prof Ella Pereira, the Computer Science department academic, support staff and my GTA colleagues for their immense support and invaluable guidance towards the research.

# Declaration of Originality

I, Bappaditya Debnath, hereby declare that this thesis is my own work and reports my own original research. I confirm that the thesis has not been submitted for another degree or professional qualification to this or any other university or learning institution. The experiments conducted is entirely my own work and the results reported are true to best of my knowledge. Due references have been provided on all supporting literature and resources.

# List of Publications

**Contribution Towards PhD**

1. B. Debnath, M. O'Brien, Swagat Kumar and A. Behera. Attention-driven Body Pose Encoding for Human Activity Recognition. 25th IEEE International Conference on Pattern Recognition (ICPR). 2020.

2. B. Debnath, M. Yamaguchi, M. O'Brien and A. Behera. Adapting MobileNets for mobile based upper body pose estimation. In 15th IEEE International Conference on Advanced Video and Signal-based Surveillance (AVSS). 2018. https://doi.org/10.1109/AVSS.2018.8639378

3 B. Debnath, Swagat Kumar, M. O'Brien and A. Behera (Under review). A step towards automated functional assessment of activities of daily living, IEEE International Conference on Intelligent Robots and Systems (IROS). 2021.

4. B. Debnath, Swagat Kumar, M. O'Brien and A. Behera. Attentional Learn-able Pooling for Human Activity Recognition in Videos, IEEE International Conference on Robotics and Automation (ICRA). 2021.

5. B. Debnath, M. Yamaguchi, M. O'Brien and A. Behera. (Accepted). A review of computer vision approaches for physical rehabilitation and monitoring. Springer, Multimedia Systems. 2021.

**Doctoral Consortium**

1. IEEE/CVF Workshop for Applications of Computer Vision (WACV). 2021.

2. IEEE International Conference on Face and Gesture Recognition (FG). 2019.

**Other Contributions**

1. A. Behera, Z. Wharton, A. Kiedel and B. Debnath. Deep CNN, Body Pose and Body-Object Interaction Features for Drivers' Activity Monitoring. IEEE Transactions on Intelligent Transportation Systems. October 2020.

2. A. Behera, A. Kiedel and B. Debnath. Context-driven Multi-stream LSTM (M-LSTM) for Recognizing Fine-Grained Activity of Drivers. In 40th German Conference on Pattern Recognition (GCPR) 2018, pp. 298-314. ISSN 0302-9743 DOI `https://doi.org/10.1007/978-3-030-12939-2_21`.

3. Z. Wharton, E. Thomas and A. Behera. A Vision-based Transfer Learning Approach for Recognizing Behavioural Symptoms in People with Dementia. In 15th IEEE AVSS 2018. `https://doi.org/10.1109/AVSS.2018.8639371`.

# Abstract

This thesis intends to contribute towards Computer Vision (CV)-based functional assessment of physically impaired persons involving Activities of Daily Living (ADL). Patients rehabilitating from conditions like stroke, spinal cord injury, Parkinson's disease, among other symptoms experience difficulty in physical movements which limit their functional ability and independence. Such patients usually undergo physical rehabilitation programs which require constant monitoring of their (ADL) to record their progress. This monitoring is currently done by Healthcare professionals which is not only labour intensive and expensive but also error-prone. The study aims to address this problem by proposing CV-based methods for detecting ADL which can be used for functional assessment of patients automatically from recorded videos. This has the potential of lessening the labour-intensive manual annotation and continuous human observation, resulting in reduction of the overall cost of rehabilitation for such patients. While the current CV literature is replete with methods for human activity or ADL recognition, there are very few that aim to detect impairment-specific versions of ADL executed or exhibited by physically impaired persons. A part of the problem lies in the unavailability of labelled datasets for such activities making it difficult for researchers to develop the necessary methods to detect and recognise them. In recent years, the field of CV has seen increasing use of Deep Learning (DL) methods for ADL recognition. However, DL-based models are almost exclusively data-driven and require very large datasets often containing thousands of human activity videos to successfully train and validate. The current study attempts to address this issue by developing and contributing a novel multi-label dataset that includes labelled videos of several categories of normal and impairment-specific executions of ADL similar to what is exhibited by normal persons and physically impaired persons, respectively. This dataset has been developed under the guidance of an Occupational Therapist providing the necessary credibility to the entire exercise. This is an inter-disciplinary research involving CV, Artificial Intelligence and Health and Social Care.

One of the key focus of this thesis is to contribute towards the advancement of research in DL. To this end, the thesis presents three novel human activity recognition models based on DL. The first model uses an intelligent or learn-able pooling method based on Fisher Vector (FV) to propose a better alternative to the standard statistical pooling method known as Global Average Pooling (GAP). In this model, FV with activity-aware pooling method is integrated within the DL model to semantically cluster the structural information contained in Attention-focused hidden LSTM states in a novel manner. It leads the network to pool more relevant information in contrast to normally used statistical pooling methods. The model achieves better performance than the state-of-the-art video-based models. The second activity recognition model introduces a novel 3D human body-pose encoding method. The body-pose encoding algorithm learns the spatial arrangement between various body joints to present an enriched pose information to the network for improved performance. The algorithm also encodes the frame-wise positions of body joints and presents a temporally enriched representation for each joint, individually. The pose encoding algorithm coupled with an Attention mechanism is presented as a part of combined video and pose-based activity recognition model that achieves state-of-the-art results on three challenging benchmark datasets. The third is a pure human body pose-based lightweight DL model based on Temporal Convolution Networks. The

spatial-temporal two-stream model takes advantage of the pose encoding algorithm and the learnable pooling method introduced earlier to impact the model performance, positively. The model is not only able to recognise an ADL, but also discriminate between the normal and different physical impairment-specific variations of the same ADL when evaluated on the multi-label dataset. Thus, it fulfills the main research aim. To the best of my knowledge, this is an unique inter-disciplinary research that attempts to recognise physical impairment-specific ADL through multi-label video analysis and recognition. In addition to the three activity recognition models, the thesis presents a mobile-based DL approach for human pose estimation. The model introduces a novel Split-Stream architecture as an alternative to the standard GAP method present towards the end of many DL models. The thesis also presents a critical review of existing research on CV-based rehabilitation and assessment. The review proposes its own taxonomy and analyses articles from a CV perspective compared to other reviews that mainly focus on the clinical perspective. The literature review, the mobile-based human body-pose estimation model, the multi-label dataset and the three human activity recognition models are the major contributions of this inter-disciplinary research.

# Contents

# List of Figures

# List of Tables

# List of Abbreviations

**ADL** Activities of Daily Living. 12, 24–30, 32–34, 36, 38, 63, 69, 74, 76, 90, 105–111, 113, 114, 118, 120, 122–126, 143, 144, 158, 159, 174–176, 178–181, 184

**AE** Auto Encoder. 65, 67

**AI** Artificial Intelligence. 24, 26, 30, 61, 67, 76, 83, 105–107, 123–126, 178, 179, 181, 184, 185

**AIMS** Abnormal Involuntary Movement Scale. 66, 67

**ANN** Artificial Neural Network. 32, 43, 50, 54, 56, 60–63, 76–78, 80, 85

**ANOVA** Analysis of variance. 43, 50, 57, 73

**AVSS** Advanced Video and Signals-based Surveillance. 34, 104, 180

**BCI** Brain-Computer Interface. 50, 51, 70

**BI** Barthel Index. 25

**Bi-LSTM** Bi-Directional Long Short-Term Memory Network. 34, 80, 126, 127, 131, 132, 134, 135, 140, 146, 148, 150, 154, 163, 182

**BN** Batch Normalization. 161

**BoVW** Bag of Visual Words. 86

**BoW** Bag of Words. 72

**CNN** Convolutional Neural Network. 31, 33–35, 41, 61–63, 65, 67, 68, 74, 76, 77, 79–82, 85–89, 91, 92, 98, 103, 126–128, 131, 132, 136, 138, 139, 142, 143, 145, 146, 149, 151, 160, 182

**CPM** Convolutional Pose Machines. 67

**CRF** Conditional Random Field. 50, 73

**CV** Computer Vision. 12, 18, 24, 26–32, 34–43, 45, 46, 49, 50, 53, 56, 69, 70, 74, 76, 83, 85, 86, 92, 93, 101, 104, 106, 118, 123–127, 143, 158, 167, 176, 178–181, 184, 185

**DBI** Davies Bouldin Index. 169–172, 177

**DCGAN** Deep Convolutional Generative Adversarial Network. 61

**DL** Deep Learning. 24, 26, 28–34, 36, 38, 46, 47, 50, 53, 56, 61, 63, 65, 67–70, 72, 74–77, 80, 82, 83, 85, 86, 91–93, 96, 105, 118, 123, 125, 126, 130, 176, 179–185

**DMW** Dynamic Manifold Warping. 55, 56

**DPC** Depth-wise and Point-wise Separable Convolution. 94, 95, 97–99

**DTW** Dynamic Time Warping. 40, 47, 51, 55–57, 60, 62, 65–68, 72, 73, 85, 185

**EMD** Earth Mover Distance. 67

**FC** Fully Connected. 31, 33, 63, 76, 77, 79, 80, 85, 95, 98, 128, 131, 139, 140, 142, 150, 154, 161–163, 182

**FIM** Fisher Information Matrix. 129, 130

**FLOps** Floating Point Operations. 175

**FMA** Fugl-Meyer Assessment. 25, 41, 43, 64, 66

**FV** Fisher Vector. 14, 19, 31, 33, 34, 80, 88, 126–135, 139, 140, 142, 143, 159, 160, 162–164, 168, 169, 171–173, 175, 177, 178, 181, 182, 184

**GAN** Generative Adversarial Network. 61, 72, 75, 83, 185

**GAP** Global Average Pooling. 31, 33, 34, 79, 95, 98, 99, 103, 128, 131, 132, 139, 140, 142, 150, 154, 160–163, 169, 170, 182

**GBM** Gradient Boosting Machines. 62, 63

**GMM** Gaussian Mixture Model. 65, 67, 80, 129, 130, 135

**GNN** Graph Neural Networks. 67, 74, 89

**GP-LVM** Gaussian Process Based Latent Variable Model. 51, 57

**GPU** Graphics Processing Unit. 28, 33, 77, 82, 93, 104, 136, 160, 165, 167, 180, 183

**GRBM** Gaussian restricted boltzmann machines. 60, 62

**HMM** Hidden Markov Model. 40, 47, 55–57, 60, 62, 66–68, 72, 73, 83–85

**HOF** Histogram of Optical Flow. 85, 86

**HoG** Histogram of Oriented Gradients. 72, 82, 85

**HSMM** Hidden Semi Markov Model. 12, 62, 66, 73

**ICPR** International Conference on Pattern Recognition. 34, 158, 181

**ICRA** International Conference on Robotics and Automation. 34, 143, 181

**IROS** Intelligent Robots and Systems. 35, 125, 178, 181

**KLD** Kullback Leibler Divergence. 73

**KNN** K-Nearest Neighbour. 49, 54, 62, 84, 85

**LDA** Linear Discriminant Analysis. 57, 62

**LID** Levodopa Induced Dyskinesia. 66, 67

**LMC** Leap Motion Controller. 47, 70

**LSTM** Long Short-Term Memory. 19, 31, 34, 35, 41, 47, 57, 61, 63, 68, 74, 76, 78, 79, 81, 82, 85–89, 91, 127, 133, 134, 138–140, 142, 146, 151, 153, 160, 182

**MBH** Motion Boundary Histogram. 85

**MDDTW** Multiple Dimensional Dynamic Time Warping. 62, 73

**MHI** Motion History Image. 65

**ML** Machine Learning. 40, 53, 55, 56, 60–64

**MLP** Multi-layer Perceptron. 54

**MSE** Mean Square Error. 69, 99–101

**NLP** Natural Language Processing. 31

**ORB** Oriented FAST and Rotated BRIEF. 72

**PAA** Piece-Wise Aggregation Approximation. 65

**PCA** Principal Component Analysis. 51, 57, 62, 67, 84, 85

**PCK** Percentage of Correct Keypoints. 101–103

**PD** Parkinson's Disease. 24, 37, 54, 55, 65–67

**PIS** Participant Information Sheet. 118

**POMDP** Partially Observable Markov's Decision Process. 47

**R-CNN** Regional Proposal Network CNN. 65, 82

**RANSAC** Random Sample Consensus. 40, 46, 55, 56

**RBF** Radial Basis Function. 54

**ReLU** Rectifier Linear Unit. 161

**RF** Random Forest. 60, 62, 63, 66–68, 85

**RGAN** Recurrent Generative Adversarial Network. 61

**RNN** Recurrent Neural Network. 65, 78–80, 89, 146, 153

**SAU** Skeletal Aaction Unit. 55

**SEU** Spatial Encoding Unit. 87, 89, 146–148, 150, 154, 155, 158–160, 162–164, 168, 169, 175, 177, 178, 182, 185

**SGD** Stochastic Gradient Descent. 100, 150, 165

**SIFT** Scale-Invariant Feature Transform. 50, 72, 85, 86, 96

**SSDTW** Sub-Sequence Dynamic Time Warping. 55, 57

**SURF** Speeded-Up Robust Features. 50, 72

**SVM** Support Vector Machines. 12, 35, 43, 50, 54–56, 60–63, 65–68, 73, 84, 85

**SVR** Support Vector Regressor. 68

**TASS** Temporal Alignment Spatial Summarisation. 55–57, 72

**TCD** Temporal Commonality Discovery. 55, 57

**TCN** Temporal Convolutional Networks. 31, 47, 67, 74, 76, 79, 83, 85, 88, 91, 146, 160–162, 173, 177, 185

**TEU** Temporal Encoding Unit. 89, 146, 148, 150, 154, 155, 158–160, 162–164, 168, 169, 175, 177, 178, 182, 185

**UKF** Unscented Kalman Filter. 47

**UPDRS** Unified Parkinson's Disease Rating Scale. 41, 64, 65, 106

**VLAD** Vector of Laterally Aggregated Descriptors. 80, 85

# Chapter 1

# Introduction

## 1.1 Introduction

Human activity understanding has received significant attention from the Computer Vision (CV) community (Li et al., 2004; Li et al., 2008; Liu et al., 2019a; Cao et al., 2018; Vakanski et al., 2018; Baradel et al., 2018b; Shahroudy et al., 2015) finding wide range of applications including, but not limited to Sports (Thomas et al., 2017), Robotics (Coşar; Bellotto, 2020), Intelligent Transportation (Xing et al., 2019; Behera et al., 2018) and Healthcare (Esteva et al., 2019; Marco; Farinella, 2018). This research particularly aims to contribute towards the Healthcare sector where there has been an increased interest in using CV-based human motion understanding for rehabilitation and assessment of physically impaired persons (Sathyanarayana et al., 2018). Physically impaired individuals include people affected by diseases such as stroke, Parkinson's Disease (PD), injuries to their spinal cord or any part of their musculo-skeletal system. Such patients among other things, often experience problems with physical movement and balance and face difficulties while performing day to day living tasks otherwise known as the Activities of Daily Living (ADL) (Ferrucci et al., 2010). To recover, improve or avoid further loss of physical functionality, such patients need to undergo physical rehabilitation programs (Ferrucci et al., 2010). Rehabilitation involves helping patients to carry out repetitive therapeutic exercises or ADL and assessing their progress over time. These activities are usually carried out by Healthcare Professionals (Physicians, Occupational Therapists, Physiotherapists) at home or in a clinic (Ferrucci et al., 2010). The assessment part of this process could be automated using CV-based methods that can recognise the normal and impaired physical activities being carried out by patients. This will help to reduce not only the cognitive load on the caregivers, but may also minimise the overall cost of administering such therapeutic services by reducing the human involvement in the assessment process. The current study is a step towards automating the functional assessment of various ADL. The main aim of this study is to recognise different ADL and differentiate between normal and different types of impairment-specific ADL as performed by physically impaired persons (Figure 1.1). This is an interdisciplinary research involving CV, Deep Learning (DL)-based Artificial Intelligence (AI) and Health and Social Care, where a DL-based CV application aims to improve functional assessment of ADL.

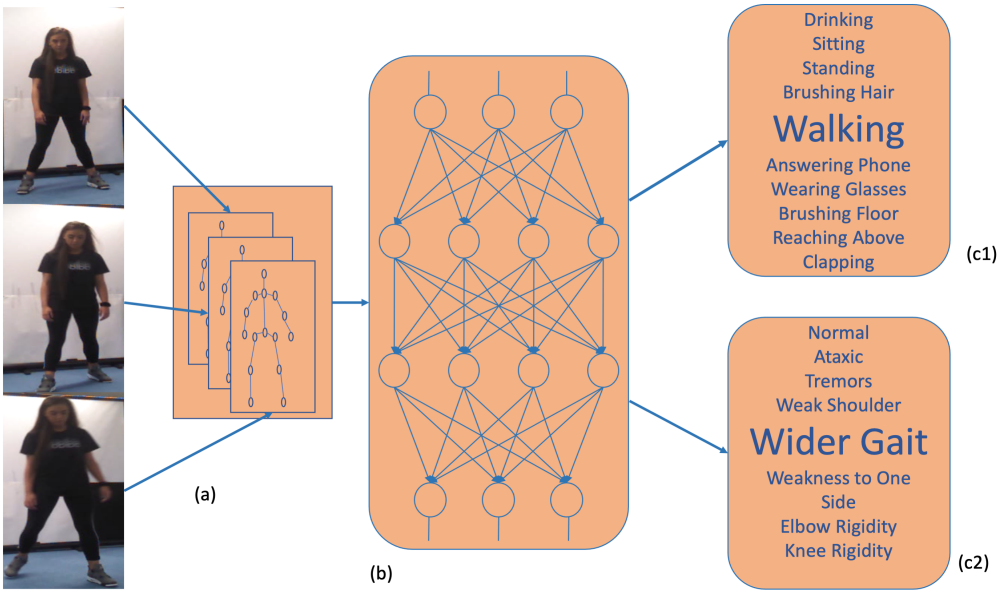



Figure 1.1: (a) Functional ADL dataset with video, depth and human body-pose information (b) DL-based model. (c) Model's prediction: c1) Activity c2) Impairment. The study explores CV-based ADL recognition, where a DL-based model aims to recognise various versions of the same ADL as performed by physically impaired persons in addition to normal ADL recognition. To achieve the aim, the study presents a functional ADL multi-label dataset where each sample has one label for 'Activity' (e.g., Walking) and another for 'Impairment' (e.g., Wider Gait). This dataset is then used to train a pose-based DL-based model for multi-label activity recognition.

### 1.1.1 Background

A functional assessment indicates a subject's ability or functional level to perform any functional work in a safe and dependable manner (Reiman; Manske, 2011). In Healthcare, there are various approaches for functionally assessing patients undergoing physical rehabilitation (Gladstone et al., 2002; Marvin; Zeltzer, 2015; Green; Young, 2001). The choice of approach depends on the type of impairment or disease involved. For example, an existing approach for assessing the mobility function of patients recovering from stroke is through Fugl-Meyer Assessment (FMA) (Gladstone et al., 2002). In FMA, patients recovering from stroke are assessed for motor function, sensory function, balance, range of motion of joints and joint pain (Gladstone et al., 2002). Another approach involves functional independence measurement while performing ADL through Barthel Index (BI) (Marvin; Zeltzer, 2015). BI uses 10 variables describing various ADL such as feeding, bathing, grooming and so on. Each variable is rated on a scale of 0 to 10, where 0 is fully dependent and 10 is fully independent. Functional assessment through ADL is widely carried out for assessing a patient's condition and various methods have been proposed to measure the same (Green; Young, 2001). Such methods often require visual observation by clinicians where the progression is recorded through pen and paper, mobile applications or a combination of both. Visual observations which rely on a clinician's experience and skill-level may suffer from errors due to subjectivity of these behavioural and clinical assessments (Mousavi; Khademi, 2014). Moreover, the assessment process requires clinicians to spend time with patients which is expensive and a major source of expenditure for both the NHS and patients. Statistics show that informal care for rehabilitation is the reason behind 27% of the whole treatment cost. In the case of stroke patients, this was around 2.42 Billion pounds

in 2016 (Stroke Association UK, 2017). The social care cost for patients suffering from spinal cord injuries or other nervous system related diseases was around 8.2 Billion pounds in the UK in 2016 (Comptroller and Auditor General, 2015). In addition to that, there are some issues pertaining to long and tedious interaction between patients and Healthcare professionals. According to a study by the Stroke Association about 48% of stroke survivors and their carers reported problems caused by either poor or non-existent co-working between care providers (Stroke Association UK, 2017). The domain of CV-based rehabilitation and assessment aims to address some of the above-mentioned issues by automating aspects of physical rehabilitation and assessment (Sathyanarayana et al., 2018; Da Gama et al., 2015b).

In recent times, CV-based research has been largely data-driven, mainly influenced by progress in the field of AI, especially DL (Voulodimos et al., 2018). Some of the best performing models in the area of image recognition (Tan; Le, 2019; Szegedy et al., 2017), human activity recognition (Baradel et al., 2018b; Shahroudy et al., 2017), human pose estimation (Cao et al., 2018) and so on have relied on DL-based architecture to achieve state-of-the-art results. Further, progress in CV-based Healthcare has also been significantly influenced by DL (Esteva et al., 2019). However, relevant literature reviewed by Sathyanarayana et al. (2018), Da Gama et al. (2015b), Mousavi; Khademi (2014) and Webster; Celik (2014) show that CV-based rehabilitation and assessment research is yet to fully explore and exploit DL. DL-based models (Tan; Le, 2019; Baradel et al., 2018b; Shahroudy et al., 2017; Cao et al., 2018) are at the forefront of research in their respective domains and research in CV-based rehabilitation and assessment can potentially benefit from the same. One of the reasons for the success of today's DL-based model is the availability of large datasets (Goodfellow et al., 2016). However, for CV-based assessment and rehabilitation authors have largely used their own small in-house datasets to conduct their research (Sathyanarayana et al., 2018; Da Gama et al., 2015b; Mousavi; Khademi, 2014; Webster; Celik, 2014). This could be one reason why research in this area has seen relatively less use of DL-based models. To fully realise the potential of CV towards rehabilitation and assessment of physically impaired patients, researchers need to explore the use of DL-based methods with the support of large publicly available datasets. This study attempts to advance the domain of CV-based rehabilitation and assessment by exploring the use of DL for functional assessment of ADL.

### 1.1.2 Motivation

The human body exhibits wide variety and range of motions which might differ from normal patterns in the case of physically impaired individuals. This leads to a wide range of impairments and it is very difficult to capture or model every abnormality in any single application. Researchers in CV-based rehabilitation and assessment have attempted to monitor patients for specific impairments, which are often limited to assessing single limb movements or very specific types of repetitive movement (Sucar et al., 2008a; Paiement et al., 2014). For example, to assess shoulder movement, shoulder flexion and abduction angle has been used by Da Gama et al. (2012). Other scenarios include repetitive movements of a body part such as the leg, to assess a specific abnormality such as abnormal gait (Pei et al., 2016). In such cases comparing joint angle trajectory of one or a few body joints may be enough for automated assessment. On the other hand, ADL is a more complex task which

requires a series of body movements involving multiple body parts. ADL are neither specific nor repetitive in nature making automated assessment more challenging as compared to gesture/posture recognition, joint angle trajectory comparison and so on. For physically impaired persons, ADL performance is widely used to assess a patient's functional ability or independence (Green; Young, 2001). This study aims to improve automated functional assessment of ADL, by recognising an ADL and differentiating between a normal and various impairment-specific variants of the same ADL. The variations in executing these ADL tasks arises from conditions such as ataxia, tremors and other conditions that a physically impaired person may have. Discriminating ADL as performed by persons with such conditions from a normally executed ADL will be a steppingstone towards automated functional assessment involving ADL. The CV community has extensively explored human activity recognition (Ke et al., 2013) including regular ADL recognition. This has been partly driven by the increasing availability of large publicly available datasets (Shahroudy et al., 2016; Wang et al., 2012). Such datasets are available in different modalities such as monocular RGB videos, depth information, human body-pose or any combination of these (Shahroudy et al., 2016; Wang et al., 2012). However, for ADL recognition for patients, datasets are currently not available for discriminating between an ADL performed by a healthy individual and the same ADL performed by a patient with one or more physical impairments. Consequently, existing human activity recognition models cannot differentiate between an ADL performed by a healthy individual versus a physically impaired individual. Motivated by the above facts, the current study presents a new dataset that consists of ADL performed by healthy individuals. It also contains the same ADL executed by healthy persons who acted like physically impaired individuals, demonstrating various physical impairments. Then, this dataset is used to train a novel CV-based multi-label activity recognition model that predicts the 'Activity' as well as the 'Impairment' associated with a given ADL sequence. The model presented here is first able to first recognise an ADL (e.g., walking, drinking etc.) and then indicate whether the ADL is a normal or one of the four impairment-specific versions of the same ADL (e.g., ataxic, tremors etc). This is different from normal human activity recognition, which only differentiates between activities such as drinking, walking and so on. Building on this motivation, the research question is posed next, followed by the formulation of the main aim and objectives to help address the main research question.

## 1.2 Research Question

The main research question is:

*How can a machine or computer recognise different activities of daily living and their variations when executed by a healthy individual versus people with different impairments?*

## 1.3 Aim and Objectives

### 1.3.1 Aim

The main aim of the research is to contribute a novel model that can not only recognise an ADL, but also discriminate the impairment-specific variations of the same ADL as executed by persons with different physical impairments in comparison to healthy individuals.

### 1.3.2 Objectives

**1** To conduct an in-depth and critical review of existing literature in CV-based physical rehabilitation and assessment.

**2** Make advancement towards lightweight human pose estimation, which could be used for mobile-based human activity recognition.

**3** Prepare a dataset that captures normal and physical impairment-specific versions of different day to day activities or ADL.

**4** Use the latest advancement in the field of DL to develop a novel ADL recognition model.

**5** Further advance the ADL recognition model to discriminate between different executions of the same ADL.

### 1.3.3 Objective formulation

- **Objective 1:** The foundation to any study starts with a review of existing literature and researchers often rely on existing reviews and surveys to gain an insight into the problem. Existing reviews and surveys in this domain mostly analyse articles from Healthcare perspective (Mousavi; Khademi, 2014; Da Gama et al., 2015a) and are yet to explore this domain from a CV perspective. Sathyanarayana et al. (2018) review articles from CV-perspective, but have not captured the latest articles after 2014. Thus, one of the objectives of this study is to conduct a review of relevant literature in CV-based rehabilitation and assessment from a CV perspective, while also capturing the latest advancements in this domain.
- **Objective 2:** Human pose estimation plays a significant role in automated assessment of physically impaired persons. Estimated body-pose is used for joint angle trajectory comparison (Exell et al., 2013), gesture recognition (Lin et al., 2013c) and so on. Authors (Mousavi; Khademi, 2014; Da Gama et al., 2015a) in this domain have largely relied on Kinect-based pose estimation whereas the current best-performing pose estimation models rely on DL-based architectures (Cao et al., 2018; Newell et al., 2016). The best performing DL-based models require high-performance Graphics Processing Unit (GPU) for inference which may not be suitable in a home or clinic environment. In such environments, lightweight applications are more suitable. Therefore, the aim is to introduce a novel lightweight pose estimation method adapted from the well-known MobileNets (Howard et al., 2017) architecture.

- **Objective 3:** CV-based assessment and rehabilitation of physically impaired persons is yet to be fully explored by the CV community. This is especially true in the context of application of DL algorithms, which has otherwise found ubiquitous use in areas such as human pose estimation and human activity recognition. Today's DL-based models are largely data-driven and one reason for the lack of interest could unavailability of publicly available large-scale datasets demonstrating patient activity. Motivated by the need to capture ADL as performed by patients, this study presents 10 different ADL with four impairment-specific and a normal execution of each ADL.
- **Objective 4:** In this work, the main investigation area is CV-based human activity recognition which includes ADL recognition. This has been extensively explored by the CV community. For recognising normal ADL as performed by healthy individuals, authors have evaluated their model performance with standard benchmark datasets that are publicly available. This helps to establish the novelties in the context of broader literature and helps to be accepted in peer-reviewed publications. Thus, this study aims to introduce novel human activity recognition models that are trained and evaluated on standard and well-known human activity recognition datasets.
- **Objective 5:** After preparing the proposed dataset, the main aim of this study is to present a DL-based model that is able to recognise ADL and discriminate between healthy and four different impairment-specific versions of the same ADL. To this end, a novel multi-label activity recognition model is prepared that take advantage of the methods introduced in the general human activity recognition models (Objective 4). Novelties from the human activity recognition models introduced earlier are used to enhance this model's performance. This objective completes the main aim of this study and answers the research question posed in the previous section (Sec. 1.2).

## 1.4 Research Focus

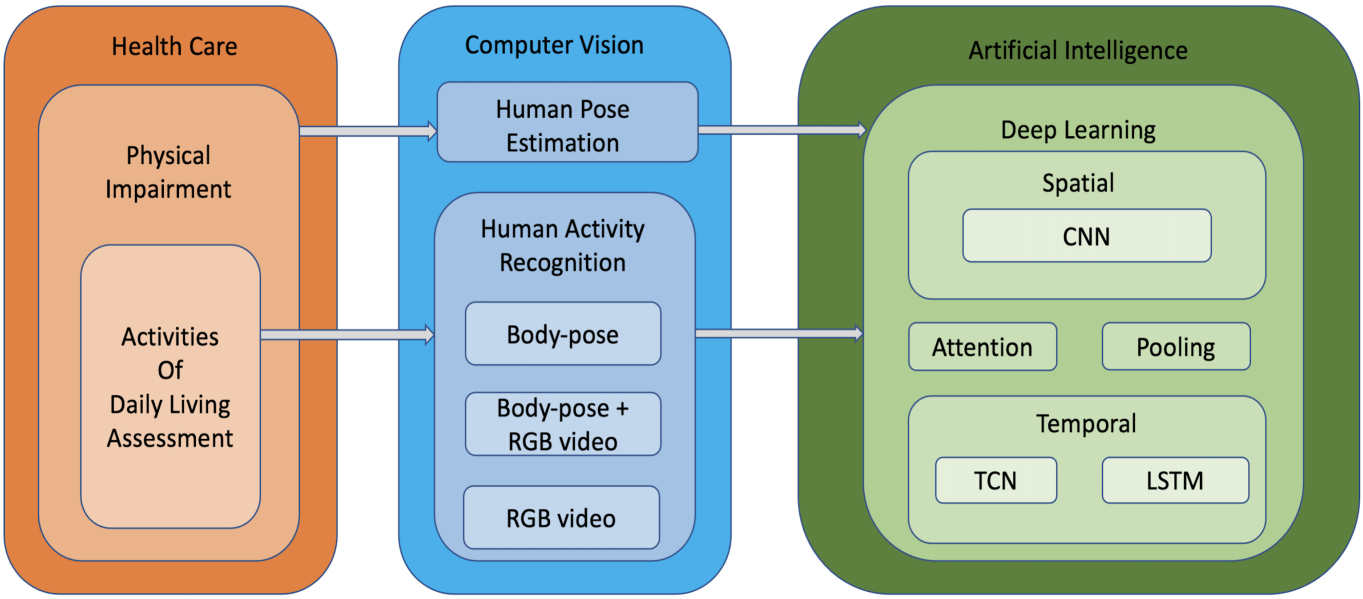



Figure 1.2: The study is an interdisciplinary research involving Health and Social Care, CV and AI. The problem of ADL assessment for physically impaired persons is from Healthcare, while the proposed solution is in the area of CV-based human activity recognition. The core focus and the novel explorations of this study is the area of DL, where CNN, TCN and LSTM are used to design the DL-based models. The study further explores 'Attention' mechanism and 'Pooling' techniques often used in DL models

### 1.4.1 Area of Research

The study is an interdisciplinary research project involving Computer Science, Artificial Intelligence and Health and Social Care. The proposed solution is built around CV with the use of AI. CV is an area of Computer Science that deals with automatic extraction, analysis and comprehension of image or video (Forsyth; Ponce, 2012). This study aims to address the problem of healthy and physical impairment-specific ADL recognition, which falls in the broader area of human activity recognition. Contemporary research in CV including human activity recognition is largely data driven (Voulodimos et al., 2018) and existing reviews show increasing use of DL-based algorithms for activity recognition (Vrigkas et al., 2015). As shown in Figure 1.2, DL is a sub-area within the broader area of AI and it mainly consists of deep neural networks (Goodfellow et al., 2016). In recent literature, the problem of human activity recognition has been addressed as a multi-class classification task where often a DL-based algorithm is trained on a labelled dataset in a supervised manner (Vrigkas et al., 2015). However, in addition to recognising ADL, this study aims to recognise the difference between regular and the physical impairment-specific variations of the same ADL. This means unlike regular activity recognition, where there is a single label (e.g., walking, drinking etc.), there are two labels. The first is 'Activity' label (e.g., drinking, walking etc.) and the second is 'Impairment' label indicating the variations within the activity (e.g., normal, ataxia, tremors etc.). This type of classification where there is more than one label for each input sample is called multi-class, multi-label classification. To achieve the same, the study introduces a multi-class, multi-labelled dataset that contains two labels for each data sample. Such a dataset can be captured with

a visual sensor such as an RGB camera, but this is not sufficient for capturing the human body pose in 3D. To capture body-pose in 3D, researchers have often used devices like Kinect which uses RGB + depth information to estimate human body pose (Mousavi; Khademi, 2014). Thus, in this study the data has been captured through Kinect, which provides RGB data, depth data and human body-pose information in 3D.

### 1.4.2 Core Focus

In recent years, the CV community has increasingly relied on DL-based models for human body-pose estimation (Cao et al., 2018; Newell et al., 2016) and human activity recognition (Shahroudy et al., 2017; Baradel et al., 2018a), owing to the superior performance of such models as compared to classic CV techniques (e.g., engineered features with discriminative models). Therefore, the core focus and novel explorations of this study are centered around DL-based human activity recognition. As shown in Figure 1.2, the study involves several key areas of DL. First, it adapts the well-known Convolutional Neural Network (CNN) architectures for pose estimation and activity recognition. CNNs are well-known for spatial processing of images and thus, these networks are well-suited for the purpose of this study. However, architectures such as Inception-ResNet-V2 (Szegedy et al., 2017), MobileNets (Howard et al., 2017) have been designed for object detection and recognition. The novelty of this study focuses on adopting these networks for pose estimation or activity recognition. Moreover, human activity recognition involves temporal data such as RGB videos or sequences of human body-pose. To efficiently capture meaningful information contained in temporal sequences, the study explores the use of Long Short-Term Memory (LSTM) and Temporal Convolutional Networks (TCN), which are well-suited for temporal processing. The human activity recognition models presented in this study combines CNN, LSTM and TCN to present novel models that give us state-of-the-art results on publicly available benchmark datasets.

Apart from these broader areas (CNN, LSTM, TCN), the study also investigates two well-known techniques known as 'Pooling' and 'Attention' mechanism. Many well-known CNN architectures use a standard Global Average Pooling (GAP) + Fully Connected (FC) layer towards the end of the network (Howard et al., 2017; Szegedy et al., 2017; He et al., 2016). But 'Pooling' using statistical methods (e.g., GAP), do not consider the temporal and other structural information captured in the network. Instead, statistical methods simply take the average value (GAP), the max-value (max-pooling) or any other statistic. This study investigates two novel alternatives to statistical pooling for enhancing the model performance. The first method is presented in the pose-estimation model and is called 'Split-Stream' architecture. This method aims to reduce the number of parameters in the FC layer which reduces over-fitting and improves the network performance. The second is a Fisher Vector (FV)-based (Perronnin; Dance, 2007) learn-able pooling method that aims for intelligent pooling instead of statistical pooling through GAP layer. Apart from 'Pooling' the study explores 'Attention' mechanisms which have greatly enhanced the performance of DL networks used for Natural Language Processing (NLP) (Vaswani et al., 2017; Zhang et al., 2018; Xu et al., 2015a; Cho et al., 2015). In NLP text data is used (Vaswani et al., 2017; Zhang et al., 2018), which are sequential in nature and is similar to video data or body-pose sequence used for activity recognition. Inspired by NLP, many authors have adapted 'Attention' mechanisms for human activity recognition

(Song et al., 2017; Sharma et al., 2016; Baradel et al., 2017). The current study explores the use of Sequential Self-Attention (Zhang et al., 2018) and Multi-Head Attention (Vaswani et al., 2017) mechanisms for the proposed activity recognition models.

## 1.5 Chapters Outline

### 1.5.1 Chapter 2: Literature review: CV-based Physical Rehabilitation and Assessment

This Chapter is motivated by the lack of reviews and surveys in the domain of CV-based rehabilitation and assessment from a CV perspective. It presents literature in this area from CV perspective where the focus is on the intelligent processing for feature extraction and comparison algorithms employed for automated assessment of physically impaired persons. The study presents its own taxonomy necessitated by the lack of review articles in this domain from CV perspective. Each article reviewed in this Chapter is tabulated to highlight the nature of raw data, feature extraction techniques and the feature comparison algorithms employed to assess patient movement. This is followed by an analysis section, which highlights the algorithms used, their pros and cons and suggests meaningful alternatives. The Chapter also highlights the lack of large-scale publicly available datasets, lightweight DL-based pose estimation and physical impairment-specific ADL recognition methods in the current literature. These findings helped to formulate the aim and objectives of the current study.

### 1.5.2 Chapter 3: Literature review: Deep Learning

This Chapter summarises the literature on DL that has been studied in preparation for the Artificial Neural Network (ANN)-based models presented in this thesis. It first discusses DL-based research necessary to understand the basics of models introduced in the current study. This includes a study of deep spatial processing networks for image data and temporal processing networks for sequential data such as human body-pose sequence or video data frames. The section includes literature on 'Attention' mechanism and 'Pooling' mechanism in DL networks as this study extensively explores these concepts (Figure 1.2). This is followed by a review of CV-based human pose estimation and human activity recognition methods. The discussion on human activity recognition is split into RGB video-based and pose-based modalities which are the basis of models presented in Chapter 6 to 8.

### 1.5.3 Chapter 4: Lightweight human pose estimation

Chapter 2 shows that accurate human body-pose estimation is essential for automated CV-based assessment of physically impaired persons. Unlike other areas where DL-based pose estimation has been widely used (Cao et al., 2018; Newell et al., 2016), authors have largely relied on Kinect for patient assessment (Sathyanarayana et al., 2018). This area needs to move to DL-based patient assessment for higher model performance. But, for home or clinic-based pose-estimation powerful

GPUs are often infeasible and lightweight pose estimation is required. In this Chapter, the well-known mobile-based DL architecture MobileNets (Howard et al., 2017) is adapted for enhancing human pose estimation accuracy.

### 1.5.4 Chapter 5: Functional Activity Recognition Dataset

This Chapter addresses the third objective of this project, which is to prepare a dataset that presents physical impairment-specific versions of various ADL. The dataset presented in this Chapter includes 10 common ADL filmed with 10 subjects. The dataset contains four impaired and one normal version for each ADL. The Chapter describes the formulation of each ADL and its impairment-specific variants included in the dataset in details. It also presents data modalities and format that any potential user will find useful for using the dataset.

### 1.5.5 Chapter 6: Human Activity Recognition: Model 1

This Chapter addresses the fourth objective of this study and presents an ADL recognition model that is purely based on monocular RGB videos. Continuing from the pose-estimation model, this Chapter also explores a better alternative to the standard GAP + FC layer. The approach suggests the use of learn-able pooling methods based on FV instead of the standard GAP + FC layers towards the end of a CNN. The experiments presented prove the efficacy of the FV-based method, enabling the use of this method in the multi-label activity recognition method presented in Chapter 8.

### 1.5.6 Chapter 7: Human Activity Recognition: Model 2

In this Chapter, a combined RGB and human body pose-based activity recognition method is presented. The model explores a novel spatial and temporal encoding method for encoding the structural information contained in the human body-pose information. These encodings enhance the performance of the pose-stream and contribute towards the overall performance of the model. The novel encoding method introduced in this Chapter helps to improve the performance of impaired ADL recognition method presented in the Chapter outlined next.

### 1.5.7 Chapter 8: Functional Activity Recognition

This Chapter presents a multi-label functional activity recognition model, which is able to recognise the difference between a normal ADL and various impairment-specific versions of the same ADL. The model presented in this Chapter takes advantage of the findings of the previous two chapters (Chapters 6 and 7) to present a novel two-stream pose-based architecture. The pose-based model is based on TCN-ResNet (Kim; Reiter, 2017). Experimental evaluations demonstrate that the novel adaptations comprehensibly out-perform the original TCN-ResNet model.

## 1.6 Contributions

This section briefly describes the intended contribution to knowledge of this study while Chapter 9 describes the same in details. Each of the five objectives of this project has distinct intended contribution to knowledge which are as follows:

**Contribution 1:** To the best of my knowledge, this is the only literature review that addresses recent advances in the domain of CV-based rehabilitation and assessment of physically impaired persons from a CV perspective. The study proposes its own taxonomy and tabulates articles that highlight CV-based data and features used along with subsequent application of learning algorithms for comparative analysis. The review has been accepted for publication in *Springer Multimedia Systems* after three reviews.

**Contribution 2:** A novel DL-based model has been prepared where the well-known MobileNets (Howard et al., 2017) has been adapted to an hourglass-like network for a lightweight pose estimation model. It also introduces a novel 'Split-Stream' architecture which enhances the model performance. This model was presented and published in the $15^{th}$ *IEEE International Conference on Advanced Video and Signals-based Surveillance (AVSS), 2018.*

**Contribution 3:** The study presents a new dataset that illustrates the difference between an ADL performed by healthy individuals and the impairment-specific variations of the same ADL as performed by persons with different physical impairments. The size of the proposed multi-modal dataset with 5685 videos is well-suited to train contemporary data-driven DL-based models. To benefit future research in this direction, the dataset will be made publicly available upon the completion of this study.

**Contribution 4:** The study presents two new human activity recognition models. The first is a purely RGB video-based model that introduces a novel FV-based learn-able pooling mechanism. This is an alternative for the GAP layer present towards the end of many state-of-the-art CNN models. The FV mechanism semantically clusters the hidden states of an 'Attention'-focused Bi-Directional Long Short-Term Memory Network (Bi-LSTM). To the best of my knowledge, this is the first model that exploits the structural information contained in hidden LSTM states. The model achieves stat-of-the-art results in RGB modality on the well-known NTU RGBD (Shahroudy et al., 2016) and the MSR 3D (Wang et al., 2012) dataset. This model has been accepted for presentation at the *IEEE International Conference on Robotics and Automation (ICRA), 2021.*

**Contribution 5:** The second human activity recognition model demonstrates effective combination of RGB and human body-pose data. It introduces a novel joint position encoding algorithm. Together with 'Attention' mechanism, the joint position encoding algorithm successfully enhance the model performance to give state-of-the-art results on three well-known activity recognition datasets. This model has been published in the $25^{th}$ *IEEE International Conference on Pattern Recognition (ICPR), 2020.*

**Contribution 6:** A novel lightweight pose-based model is introduced which when trained on the multi-label dataset (Objective 3) is able to discriminate between a normal and four different impairment-specific versions of the same ADL. The model takes advantage of the joint position-encoding algorithm

and the learn-able pooling method introduced earlier to enhance the performance comprehensively. Along with the dataset (Contribution 3), the multi-label activity recognition model has been submitted to the *IEEE International Conference on Intelligent Robots and Systems (IROS), 2021.*

## 1.7 Other Contributions

In addition to the publications mentioned, the concept of the dataset and the novel pose-encoding method was selected for presentation at the IEEE International Conference on Face and Gesture Recognition (FG) 2019 Doctoral Consortium. The consortium had provisions for travel bursary and conference fee grant. The presentation was mentored by Dr Yaser Sheik who is an Associate Professor at the Robotics Institute at Carnegie Mellon University. Recently, the activity recognition models were selected to be presented at IEEE/CVF Workshop on Applications of Computer Vision (WACV) 2021 Doctoral Consortium which included full conference fee wavier. The mentor Dr Ehsan Elhamifar, Assistant Professor at the Northeastern University appreciated the exploration of multi-label human activity recognition. During this research, I have also made contributions to other research in the department as listed in the 'List of publications' section. The first is 'A Vision-based Transfer Learning Approach for Recognising Behavioral Symptoms in People with Dementia' which proposed a novel dataset consisting of 65,082 images of people with dementia in aggressive, depressive, happy and neutral emotions. I contributed to the research by identifying appropriate images for dataset collection and also helped with the write-up. To our knowledge, this is the first dataset to represent CV-based recognition of behavioral symptoms in people with Dementia. To evaluate the dataset a model was presented which used Support Vector Machines (SVM) to classify features formed by fusing several fine-tuned state-of-the-art CNNs. Besides the dataset, the novelty of the work lies in the way the networks were fine-tuned and combined for the performance improvement. The next paper is 'Context-driven Multi-stream LSTM (M-LSTM) for Recognising Fine-Grained Activity of Drivers'. The paper is about recognising in-vehicle driver activities in intelligent vehicles. This is very useful for identifying dangerous activities such as using mobile phones and other distractions. The work proposes a novel model based on LSTM, which combines context-aware, body-pose and body-object features extracted through pre-trained CNNs. The extracted features create a multi-stream network that provides state-of-the-art performance on a large and challenging dataset. I contributed by pre-processing the dataset, setting up the experiments and fine-tuning the model. The journal article titled 'Deep CNN, Body Pose and Body-Object Interaction Features for Drivers' Activity Monitoring' also explores in-vehicle driver activities. The work proposes a Multi-stream Deep Fusion Network which combines high-level semantics with CNN features. Here, pre-trained CNN features from the scene are combined with novel body-pose and body-object interaction features and classified through a basic linear SVM classifier. Here, I was involved in the initial design of the framework for extracting CNN features and feeding them into a SVM for classification.

## 1.8 Conclusion

To summarise, the main aim of the study is to contribute towards the domain of CV-based functional assessment of ADL for physically impaired persons. To this end, the project first prepares a dataset that presents physical impairment-specific versions of different ADL in addition to normally performed ADL. The dataset is then used to train a new DL-based model to discriminate a perfectly executed ADL from impairment-specific versions of the same ADL. To the best of my knowledge, this is the first study that explores CV-based automated assessment of patient ADL in the form of multi-label human activity recognition. The next Chapter presents a literature review on CV-based rehabilitation and assessment. The review helped to find gaps in the existing literature and led to the formulation of the main aim and the objectives of this study.

# Chapter 2

# Literature Review: Vision-based Physical Rehabilitation and Assessment

## 2.1 Introduction

This Chapter addresses the first objective, which is to critically review existing literature on CV-based rehabilitation and assessment of physically impaired individuals. Physically impaired persons include people affected by diseases such as stroke, PD, injuries to their spinal cord or any part of their musculo-skeletal system which affect their normal physical movements. Patient's recovering from such impairments often undergo extensive physical rehabilitation which involves Healthcare Professionals helping patients to carry out repetitive therapeutic exercises and assessing their progress over time. The domain of CV-based rehabilitation and assessment of physically impaired persons aims to automate this rehabilitation and assessment process which is currently largely carried out by Healthcare professionals. This review includes articles from the last 20 years that is representative of the research carried out in the domain of CV-based rehabilitation and assessment of physically impaired persons.

First, the chapter discusses the characteristics of the domain of CV-based physical rehabilitation in terms of i) physical impairment data, ii ) feature extraction and representation techniques and iii) feature comparison algorithms that lead to the final assessment and feedback. This is followed by a discussion on the taxonomy that has been used to categorise the articles. The subsequent sections (Sec. 2.4 to Sec. 2.8) presents the articles according to the taxonomy where each reviewed article is tabulated to highlight the characteristics in terms of data, feature extraction and representation and feature comparison algorithms used. Each of these sections also present a critical summary of the methods used towards the end. This is followed by a section on the current publicly available datasets useful for research in this domain. The next section, (2.10) presents a critical analysis of the data, feature extraction and representation and feature comparison algorithms used by articles reviewed. The Chapter concludes with a discussion section that reviews the gaps in the literature that led to the formulation of objectives for the current research.

### 2.1.1 Motivation/Rationale

| Author | References | Journal | Comments |
|---|---|---|---|
| Zhou; Hu (2008) | 184 | Biomedical Signal Processing and Control | Highlights tracking methods |
| Webster; Celik (2014) | 96 | Journal of Neuroengineering and Rehabilitation | Focused on Kinect-based research, elderly care and stroke rehabilitation |
| Mousavi; Khademi (2014) | 105 | Journal of Medical Engineering | Focuses on Kinect-based research and highlights the impact of Kinect |
| Da Gama et al. (2015b) | 66 | Games for health journal | Focuses on Kinect-based research |
| Sathyanarayana et al. (2018) | 192 | Journal of Ambient Intelligence and Humanized Computing | Patient monitoring and algorithms |
| Ahad et al. (2019) | 79 | CVPR workshop | Action understanding for assistive Healthcare |

Table 2.1: Past reviews and surveys on vision-based physical rehabilitation and assessment

The current study aims to contribute towards the domain of CV-based automated physical rehabilitation and assessment and as explained in Chapter 1 (Sec. 1.3) Accordingly, the first step was to conduct a thorough literature review of existing research in this domain. To this end, this section first briefly describes the existing review and surveys in this domain (Table 2.1). Then it discusses the gaps in existing reviews which forms the basis of the current literature review. Zhou; Hu (2008) surveyed human motion tracking for rehabilitation which mainly focuses on various CV and sensor-based tracking systems. It further discusses home-based and robot-aided rehabilitation systems. The article does not describe algorithms used for comparative evaluation or physically impaired movement detection. Webster; Celik (2014) reviewed Kinect-based research and focused on the formulation of rehabilitation exercises for monitoring, where the authors discuss elderly care and stroke rehabilitation methods. Similarly, Da Gama et al. (2015b) also reviewed Kinect-based research where the focus was on the formulation of rehabilitation experiments, subsequent monitoring of progress and assessment of various comparison techniques. Most of the articles presented rely on basic methods such as average angle flexion, Euclidean distance, mean error, correlation coefficient etc. The authors present taxonomy in terms of evaluative, applicability, validation and improvement category and their taxonomy is based on a clinical perspective. Both Webster; Celik (2014) and Da Gama et al. (2015b) survey articles from a clinical perspective where clinical progress made by patients is the major focus. The articles (Webster; Celik, 2014; Da Gama et al., 2015b) do not highlight or analyse CV-based techniques used for feature extraction, feature representation, comparison or analysis of the data involved in assessment and rehabilitation. Sathyanarayana et al. (2018) reviewed articles from CV-perspective and highlighted CV-based algorithms for feature representation and comparison. Their taxonomy is based on clinical application and articles include areas such as ADL recognition or fall detection which does not always include physically impaired motion. In the absence of physically impaired motion data, it is not easy to assess a model's performance in physical rehabilitation scenarios. Moreover, the review does not include articles after 2014 and thus does not capture the latest advances made in this domain. As explained in Chapter 1 (Sec. 1.1.1), this project intends to contribute towards CV-based physical rehabilitation and assessment by exploring advanced DL methods which are mainly data-driven. Here, it is important to comprehend

the nature of raw data, feature extraction and representation methods and feature comparison algorithms. Thus, in the current literature study, articles have been reviewed from CV perspective. Here, the focus is on the nature of physical impairment data, the feature comparison and representation techniques and the feature comparison algorithms. Feature comparison algorithms are particularly important to this study focusing on methods for discriminating physically impaired activity from a healthy person's activity. Feature comparison algorithms are also important for assessing the extent of deviation from a healthy movement. In the next section, characteristics of this domain is described concerning the use of physical impairment data, feature comparison and representation techniques and feature comparison algorithms.

## 2.2 Domain Characteristics

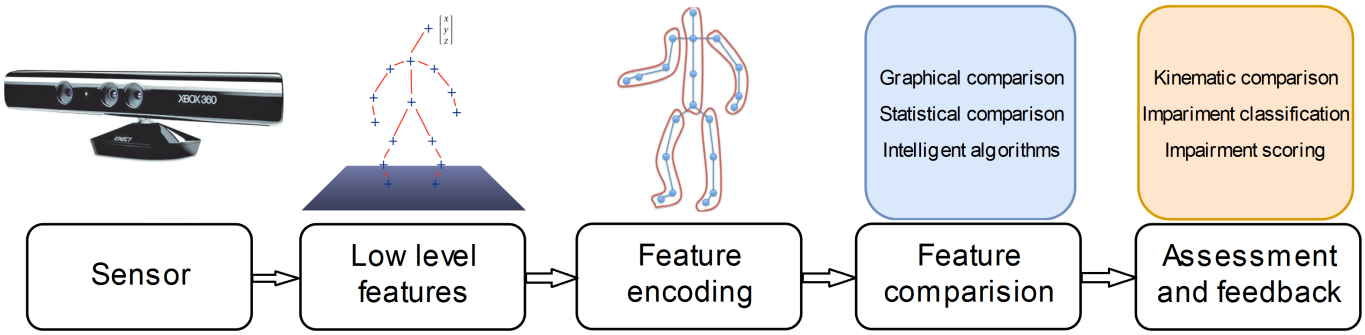



Figure 2.1: A very high-level illustration of general logical flow for a CV-based physically impaired patient assessment system

Figure 2.1 shows the general flow of research in the domain of CV-based rehabilitation and assessment. The illustration highlights important characteristics of this domain. It includes a CV-based sensor such as RGB or depth camera for sensing data. A low-level feature, such as human joint positions. A feature encoding and representation method such as a group of joint positions or a combination of human kinematic parameters. Encoded features are then compared through simple graphical and statistical techniques or through intelligent algorithms. Finally, assessment is done through kinematic parameter comparison, pose recognition, automated clinical scoring, impairment classification and other such techniques. Rehabilitation systems usually have an exercise program and provide feedback. Like any data-driven research, these characteristics can be broadly described in three major parts (primary data, feature extraction and representation and feature comparison) as discussed next.

### 2.2.1 Physical Impairment Data

For other CV-based human motion modelling areas such as human pose estimation or activity recognition, large scale datasets are publicly available (Shahroudy et al., 2016). Thus, collecting data is often outside the scope of research. However, for researches in physical impairment domain, authors have often collected their own data (Webster; Celik, 2014; Mousavi; Khademi, 2014; Da Gama et al., 2015b). Owing to the wide-range of human movements and impairments, clinicians have specific tests and exercises designed for rehabilitation and assessment of different types of physical impairments.

Therefore researchers have captured data specific to the physical impairments that they aimed to address (Webster; Celik, 2014; Mousavi; Khademi, 2014; Da Gama et al., 2015b). The need to capture data specific to impairment arises for the fact that today's CV-based models are largely data-driven (Simonyan; Zisserman, 2014; Szegedy et al., 2017; Howard et al., 2017) and learn characteristics of the data provided at the time of training the models. Thus, most authors have captured data catering to specific situations they aimed to address. Such impairment-specific datasets are often very small (Webster; Celik, 2014; Mousavi; Khademi, 2014; Da Gama et al., 2015b) and are not suitable for modern data-driven models where larger datasets containing thousands of samples are used (Shahroudy et al., 2015; Shahroudy et al., 2016). Thus, there are a very few publicly available datasets (Table 2.7) and even these are very small as compared to datasets available for other CV applications areas (e.g., image recognition, human activity recognition) (Akopyan; Khashba, 2017). Before the availability of Kinect (Su et al., 2014), researchers mostly relied on RGB video/image data (Sucar et al., 2008b; Avilés et al., 2011) and used indirect techniques like colour-ball detection (Sucar et al., 2010), silhouette detection (Leu et al., 2011) to track human body parts. With the introduction of Kinect, depth information became readily available which aided in the estimation of 3D human body-pose (Antón et al., 2013). In this study, the target abnormality, area of the body affected and the corresponding data collected has been highlighted for each reviewed article.

### 2.2.2 Feature extraction and representation

The main aim of feature extraction and representation is to select and encode raw data in a manner that improves the discriminatory power of comparative algorithms (Figure 2.1). Authors using both video data or 3D-pose estimation have aimed towards encoding various body parts or joint position for feature representation (Spasojević et al., 2015; Spasojević et al., 2017). One approach using body-pose estimation is to encode kinematic features such as joint angle trajectory, relative joint position, speed and acceleration (Spasojević et al., 2015; Spasojević et al., 2017; Natarajan et al., 2017). This is useful when a specific type of impairment is in consideration such as discriminating pathological gait, knee angle, step distance, (Paiement et al., 2014; Tao et al., 2016) and so on. Another approach is to quantify the difference between patient action and a perfect template. To achieve this, authors have used more complex representations such as temporally aligned sequence using time sequence modelling algorithms like Hidden Markov Model (HMM) (Tao et al., 2016) or Dynamic Time Warping (DTW) (Baptista et al., 2017). Articles that have used video data also have attempted to extract useful features for localisation of joint or body parts (Rivas et al., 2018; Leu et al., 2011; Cho et al., 2009). One approach is indirect tracking through colour, achieved by placing colour markers on the body or hand held devices (Rivas et al., 2018). Another approach involves extracting human body silhouette from RGB (Leu et al., 2011) or depth image (Natarajan et al., 2017). Classic CV algorithms such as histogram-based colour, texture detection (Sucar et al., 2008a; Sucar et al., 2010), Random Sample Consensus (RANSAC), morphological operations (Natarajan et al., 2017), binary silhouettes (Cho et al., 2009) have been used for feature extraction and representations from RGB or depth images. Authors have also used Machine Learning (ML) algorithms for feature extraction and dimensionality reduction (Cho et al., 2009; Leightley et al., 2013). All the articles reviewed in this study have been tabulated where the feature extraction techniques have been summarised in the

column 'Feature'. Also, the feature extraction and representation techniques have been described in the accompanying discussion to highlight the general trend.

### 2.2.3 Feature Comparison

Articles reviewed in this study have used various feature comparison techniques for CV-based assessment of physically impaired persons. Authors have used methods including, but not limited to simple graphical analysis (Leu et al., 2011), statistical analysis (Kurillo et al., 2013), time-series comparison (Zhi et al., 2018), classification (Jun et al., 2013) (Kertész, 2013) and regression (Akopyan; Khashba, 2017). For assessment of a patient's condition, the requirements vary widely from simple graphical comparison (Exell et al., 2013) to methods for automatically establishing clinical scores such as FMA (Gladstone et al., 2002), Unified Parkinson's Disease Rating Scale (UPDRS) (Rating Scales for Parkinson's Disease, 2003) etc. For simple graphical comparison, visual comparison of joint angle trajectories is enough (Pei et al., 2016). But, for other cases such as automated clinical scoring, more advanced discrimination (comparison) algorithms such as CNN-LSTM have been used (Vakanski et al., 2018). Methods such as joint angle trajectory comparison are relatively simple and may not require large datasets. On the other hand, establishing automated clinical scoring may require advanced algorithms and large datasets to work reliably. As explained earlier, obtaining large scale dataset for each type of physically impaired motion is difficult. Therefore, the main challenge in this area is to maximise the applicability of advanced algorithms with limited data. Feature comparison techniques used by each article have been highlighted in the tables under the column 'Objective'.

## 2.3 Taxonomy

The current study proposes its own taxonomy, which is necessitated due to lack of surveys in this area from a CV application point of view. As discussed earlier, any vision research has three parts 1) data collection, 2) feature extraction and representation and 3) feature comparison. The review both categorise and tabulate the articles to highlight the above-mentioned aspects. The columns headed **Target** and **Dataset** highlight the kind of impairment, area of the body affected and briefly summarises the data collected. The columns headed **Sensor/data** summarises the type of sensor(e.g., Kinect), the data type (e.g., RGB, Depth). Any non-vision hardware/sensors used along with vision-sensors have also been listed in the **Sensor/data** column. The column **Feature** highlights the feature extraction and representation algorithms. The last column under the heading **Objective** summarises the comparison method and the objective from application perspective. Most reviews on CV-based research in other areas (Ke et al., 2013; Vrigkas et al., 2015) have focused on categorising the discussion in terms of algorithms or techniques used. Articles reviewed in these reviews often have common goals such as activity recognition, pose estimation and they also use common datasets. Thus, a readily available and fair comparison between the methods used can be drawn. But, due to the wide-ranging goals of research in CV-based rehabilitation domain, authors have used very different data, feature representation and comparison methods. Hence, it is very difficult to categorise each research in terms of methods or algorithms used. Instead, the taxonomy presented is

based on end-user application as explained next. However, the discussion on each of application type has been further broken into paragraphs based on similarity of methods used. This review primarily places applications into two major categories: 'Rehabilitation' and 'Assessment'. These are further sub-categorised as:

1. Rehabilitation: Automated rehabilitation system
    (a) Virtual rehabilitation
    (b) Direct rehabilitation
2. Assessment: Point in time assessment
    (a) Comparison
    (b) Categorisation
    (c) Scoring

### 2.3.1 Rehabilitation

Articles placed in 'Rehabilitation' systems have the primary goal of providing an automated system for patients to undergo physical therapy, gesture therapy or other rehabilitation exercises. Such systems guide patients to perform their rehabilitation tasks and may be fully automated and/or Healthcare professional mediated. Research in this category usually aims to improve the patient's condition. Rehabilitation systems have been further categorised into 'Virtual Rehabilitation' (Table 2.2) and 'Direct Rehabilitation' (Table 2.3) applications. In 'Virtual Rehabilitation' applications, a patient's performance in a virtual world is assessed rather than directly assessing a patient's physical performance. This includes an avatar performing tasks in the virtual world or the use of serious games for rehabilitation. Here, subjects are required to perform activities in a virtual world through real-world movements. On the other hand, 'Direct Rehabilitation' systems (Table 2.3), guide users through a web-based interface to perform rehabilitation exercises, while their movements are directly tracked through CV-based sensor. Here, a patient's physical performance is measured instead of their avatar's performance or their ability to complete tasks in a virtual world. In both 'Virtual Rehabilitation' (Table 2.2) and 'Direct Rehabilitation' type applications, patient assessment may be inbuilt or may require Healthcare professionals.

### 2.3.2 Assessment

In articles categorised as 'Assessment' applications, the goal is to provide a point in time assessment of patients' quality of motion linked to one or more body parts (Table. 2.4, 2.7, 2.6). In such applications, assessments are carried out in a clinical or non-clinical setting but there is no rehabilitation system involved (Table. 2.4, 2.7, 2.6). In this study, 'Assessment' applications have been further categorised into three types based on how an end-user would receive the output. The first type is 'Comparison', where patient data such as kinematic parameters etc. are obtained for comparison

but there is no decisive automated scoring system available (Table 2.4). Typically, such applications, present a statistical (e.g., Analysis of variance (ANOVA) (Kurillo et al., 2013)) or simple graphical comparison (González et al., 2012) of ideal vs patient kinematic parameter trajectories. The second is 'Categorisation' type applications, where the main goal is to categorise a patient's activity into types of abnormalities or simply as normal or physically impaired movements (Table 2.5). Applications placed in this category are more decisive in nature where the main goal in CV terms is classification (2.5). In such applications, authors have used classification algorithms like SVM (Taati et al., 2012; Leightley et al., 2013), ANN (Cary et al., 2014) and so on to categorise a patient's activity decisively. The third is 'Scoring' type applications where the main goal is to provide a score to a patient's activity to assess its quality of motion (Table 2.6). Attaching a score to grade a patient's quality of movement is more decisive than 'Comparison' or 'Categorisation'. The score can be clinical scoring such as FMA (Gladstone et al., 2002) or author proposed scoring (Olesh et al., 2014). The following sections discuss the articles reviewed in this study, according to the taxonomy developed.

## 2.4 Virtual Rehabilitation

| Author | Target | Dataset | Sensor/data | Feature | Objective |
|---|---|---|---|---|---|
| Sucar et al. (2008a) | Stroke, upper limbs | 1 patient, 6 therapy sessions | RGB camera/ skin colour | Colour based hand trajectory | Gesture therapy through hand tracking |
| Sucar et al. (2010) | Stroke, upper limbs, face | 42 stroke patients, 21 therapy sessions | RGB camera, pressure gripper/ colour ball, hand pressure | Hand trajectory based on colour ball detection | Gesture therapy through hand tracking and face detection |
| Cameirão et al. (2010) | Stroke, upper limbs | 10 control subjects, 12 patients, virtual tasks therapy | RGB camera, finger tracking gloves/color marker, finger positions | lower arm and finger based kinematic parameters | ANOVA analysis based patient assessment |
| Avilés et al. (2011) | Stroke, upper limbs | 4 healthy subjects, 3 min sessions | RGB camera, pressure gripper/ colour ball, hand pressure | Hand trajectory based on colour ball detection | Assessment based on POMDP |
| Schönauer et al. (2011) | Chronic pain, whole body | 6 patients, 4 to 6 game therapy sessions | Kinetic, iotracker (Pintaric; Kaufmann, 2007)/skeleton data | FAAST (Suma et al., 2011) gesture, skeleton trajectory | Comparison of Kinect and marker based skeleton tracking |
| Kurillo et al. (2011) | Stepping in place | 12 healthy individuals, stepping exercise | RGB-D camera | segmentation algorithm based hip angle | correspondence algorithm, variance analysis |
| Da Gama et al. (2012) | General Motor rehabilitation, upper limb | 50 correct, 60 wrong movement | Kinect/ skeleton data | Shoulder and elbow angles | Correct Movement Recognition and visual guidance |
| Chang et al. (2012) | SCI, shoulder movements | 2 subjects, 1 healthy, 1 patient | Kinect, OptiTrack/ OpenNI skeleton | Right hand, elbow and shoulder trajectory | Trajectory comparison between Kinect and OptiTrack |
| Fern'ndez-Baena et al. (2012) | Knee rehabilitation | 1 subject, 9395 frames | Kinect, Vicon/ NITE skeleton | Joint angle trajectory | Comparison of ROM, ME and MER |
| Antón et al. (2013) | General rehabilitation, whole body | 5 subjects, 80 recordings | Kinect/ skeleton data | Body posture in frames | Posture recognition of 3D avatar |
| Parry et al. (2014) | Burn injury, upper limbs rehabilitation | 30 children video data | 8 camera 3d motion analysis system/skeleton data | shoulder elbow angles, elevation time | Assessment of ROM, ANOVA analysis |
| Adams et al. (2015) | Stroke, upper limb | 14 impaired arm stroke patients | Kinect/ Kinect SDK | Joint angle, rate and acceleration | VOTA metrics for clinical assessment |
| Pei et al. (2016) | Stroke, whole body rehabilitation | 3 healthy adults, 7 exercises, 20 repetitions | Kinect/ skeleton data | Joint angle comparison | Joint angle statistical analysis |
| Desai et al. (2016) | Stroke, whole body rehabilitation | 10 healthy subjects | Kinect/skeleton data | Joint angle based skeleton model | Skeleton model-based game |
| Avola et al. (2018) | PD, whole body | 20 healthy, therapy session, 3 rehab exercises | Kinect, LMC, HMD/ Skeleton data | Kinect: position, speed, angle, LMC: pinch strength, fingertip speed 3 rehab exercises involving whole body | LSTM network for providing impairment scoring |
| Yu; Xiong (2019) | General rehab, whole body | 21 old subjects, Tai Chi exercise | Kinect/ Kinect V2 Unity | Joint angle | DTW-based score |

Table 2.2: Virtual rehabilitation: systems where users perform activities in virtual world for completing rehabilitation tasks. ANOVA: Analysis of Variance, HMD: Head-Mounted Display, ME: Mean error, MER: Mean Error Relative, POMDP: Partially observed Markov's Decision Process, ROM: Range of Motion, VOTA: Virtual Occupational Therapy Assistant

The objective in virtual reality and serious games-based rehabilitation application is to provide a set of virtual tasks that will require the user to perform therapeutic gestures, rehabilitative or cognitive exercises (Table 2.2). The user's movement in the real world is tracked through devices like Kinect (Da Gama et al., 2012), or other sensing methods. The goal is to accurately reflect a user's physical movement in the virtual world, often through an avatar (Avilés et al., 2011). In 'Virtual Rehabilitation' systems, role of CV is largely limited to tracking. However, this review focuses on works with a secondary objective related to CV such as gesture recognition (Antón et al., 2013) or simple graphical comparison of trajectories (Chang et al., 2012) and so on. The discussion is split into, non-skeleton, skeleton-based and automated assessment systems.

### 2.4.1 Non-skeleton based

Figure 2.2: An example of virtual rehabilitation where performance in the virtual world is considered for assessment. Here, hand is tracked indirectly through the green ball (Sucar et al., 2010).

'Virtual rehabilitation' existed before skeleton tracking became feasible. Early research in this area used indirect methods for tracking human limb movement such as colour detection, object detection etc (Sucar et al., 2008a; Sucar et al., 2010; Cameirão et al., 2010). In 2008, Sucar et al. (2008a) used skin colour to track hands for gesture therapy where colour markers based skeleton tracking was used as a cheap alternative to inertial sensors. Sucar et al. (2010) developed rehabilitation system for hand movement of stroke patients where a green ball attached to a hand gripper was used for

tracking as shown in Figure 2.2. Similar to other researches in the category (Table 2.2), participants were required to move the concerned body part (in this case stroke affected arm) through a simulated environment and assessed through their performance in virtual environment. Stroke patients often compensate reduced hand movement through trunk and thus, trunk compensation was tracked. This was done by face detection implemented using Haar Cascade classifier (Viola; Jones, 2004). Authors also attempted to use their own skeleton tracking algorithms for rehabilitation in virtual reality (Obdržálek et al., 2012).

Non-skeleton based methods are inherently limited in ability due to unavailability of joint positions. Typically, these types of applications are able to track either one joint or a body part such as one hand (Sucar et al., 2008a). The unavailability of a joint can be sometimes compensated by using CV-based methods such as body tracking from silhouette (Leu et al., 2011; Natarajan et al., 2017). Natarajan et al. (2017) used depth information in a RANSAC-based plane fitting method to discriminate the subject plane from the background. This combined with morphological operations enabled users to select the human silhouette. Morphological operations and silhouette methods need to be tailored to the specific scenario presented (Leu et al., 2011; Natarajan et al., 2017) and are susceptible to background noises (Mellouli et al., 2017; Jamil et al., 2008; Liu; Sarkar, 2005). Thus, CV researchers have proposed various methods for image enhancement through noise removal (Mellouli et al., 2017; Jamil et al., 2008; Liu; Sarkar, 2005) to mitigate the impact of background-noise. Non-skeleton based methods (Leu et al., 2011; Natarajan et al., 2017; Sucar et al., 2008a; Sucar et al., 2010) have used their own in-house datasets. Thus, it is difficult to compare the efficacy of their proposed skeleton/body parts extraction technique with the broader CV-based literature.

### 2.4.2 Skeleton-based

With the introduction of Microsoft Kinect in 2010, skeleton tracking became readily accessible and authors have used this information to track subjects for completing tasks in virtual world (Chang et al., 2012; Da Gama et al., 2012; Fern'ndez-Baena et al., 2012). The tasks in the virtual world require a subject to attain correct therapeutic gestures to progress in the virtual world or game (Da Gama et al., 2012; Da Gama et al., 2012; Antón et al., 2013). In these approaches, joint positions and angles calculated from Kinect have been used to determine whether the subject could attain the correct posture (Da Gama et al., 2012; Da Gama et al., 2012; Antón et al., 2013). Posture/gesture recognition has been achieved by Kinect runtime (Chang et al., 2012) or other software such as OpenNI middleware (Da Gama et al., 2012). Another approach is to use Kinect-based body pose information and joint angle trajectories to calculate kinematic parameters such as range of motion, mean error in range of motion etc. (Fern'ndez-Baena et al., 2012), These parameters are then used to assess a patient's physical condition or progress in terms of rehabilitation (Fern'ndez-Baena et al., 2012; Adams et al., 2015; Pei et al., 2016; Cameirão et al., 2010). To evaluate a patient's kinematic parameters, authors have used techniques ranging from simple graphical comparison (Fern'ndez-Baena et al., 2012), statistical analysis (Pei et al., 2016) to more recent DL-based models (Avola et al., 2018).

Kinematic parameters are widely used for assessment of physically impaired persons (Webster; Celik,

2014; Mousavi; Khademi, 2014; Da Gama et al., 2015b; Sathyanarayana et al., 2018) as it can be more robust and reliable than purely visual assessment made by clinicians (Bigoni et al., 2016). Availability of skeleton tracking makes it easier to obtain kinematic parameters as compared to non-skeleton methods (Sec. 2.4.1). With skeleton-based methods authors have used kinematic parameters for tasks like joint angle comparison (Da Gama et al., 2012; Schönauer et al., 2011), statistical analysis (Pei et al., 2016; Parry et al., 2014; Fern'ndez-Baena et al., 2012), gesture recognition (Antón et al., 2013) and so on. However, graphical comparison of joint angles and use of joint angles for gesture recognition are very basic methods. Researchers have proposed more robust methods for trajectory comparison (Zhang et al., 2006) and gesture recognition (Tu et al., 2019; Palma et al., 2016a). These methods involve the use of HMM and DTW, which are well-known for their use in sequence comparison (Bishop, 2006). More recently in the wider literature, robust DL-based algorithms such as TCN and LSTM have been widely used for sequence comparison (Oord et al., 2016; Lea et al., 2017), gesture recognition (Zhang et al., 2017a) and so on. Thus, applications in this area (Table 2.2) can also potentially benefit from the DL-based algorithms.

### 2.4.3 Automated assessment

Some 'Virtual rehabilitation' systems also have integrated automated assessment (Avilés et al., 2011; Da Gama et al., 2012; Adams et al., 2015; Avola et al., 2018). The main focus in 'Virtual rehabilitation' based assessment system is to transform a patient's skeleton position obtained by Kinect into categories of correctly performed moves, correct postures etc. Avilés et al. (2011) relied entirely on objectives completed in games and used Partially Observable Markov's Decision Process (POMDP) to assess the subjects. Not taking the subject's pose into account while performing assessment may lead to inaccuracies arising from incorrectly functioning limb movement compensated by other body parts (Da Gama et al., 2012; Adams et al., 2015). Thus, Da Gama et al. (2012) relied on skeleton tracking for patient assessment to account for compensatory movements. Da Gama et al. (2012) used posture recognition for assessment but have also calculated range of motion from Kinect provided joint positions to ascertain that correct movement was maintained throughout. The Kinect skeleton is not always reliable and thus, Adams et al. (2015) have used Unscented Kalman Filter (UKF) (Wan; Van Der Merwe, 2000) to enforce realistic arm kinematics, joint angle constraints, handle noisy measurements and sensor dropouts. UKF is an advanced version of the original Kalman Filter which is more suitable for highly non-linear sequences like human skeleton trajectories (Da Gama et al., 2012). DL-based models are robust and well-suited for handling such highly non-linear data and as compared to methods like joint angle trajectory comparison (Goodfellow et al., 2016). Thus, more recently, Avola et al. (2018) have used LSTM networks to learn the impairment scores from Kinect and Leap Motion Controller (LMC) device for movements involving multiple joints of hand and palm.

## 2.5 Direct Rehabilitation Systems

| Author | Target | Dataset | Raw data | Feature | Objective |
|---|---|---|---|---|---|
| Ghali et al. (2003) | Stroke, upper limbs | 1 stroke patient | RGB camera/colour | colour based object trajectory, colour histogram | Object recognition, Event detection for text feedback to patient |
| Tao; Hu (2004) | Stroke, upper limbs | Target reaching motion | RGB Camera, Qualisys/Colour marker | Joint angles trajectory | Statistical comparison with Qualisys |
| Zariffa; Steeves (2011) | Neurological injury of palm | 10 healthy subjects, 3 types of grip | 2 RGB cameras | Hu invariant and contour signature extracted from background subtraction | Grip classification through KNN |
| Huang (2011) | Cerebral palsy and muscle atrophy, upper limbs | 4 patient subjects | Kinect/ Skeleton data | Joint angles based posture | Correct exercise count through posture recognition |
| Chang et al. (2011) | Cerebral palsy and muscle atrophy, upper limbs | 2 patients, 34 days therapy | Kinerehab Huang (2011) | Joint angles based posture | Correct exercise count through posture recognition |
| Frisoli et al. (2012) | Hemiparesis stroke, upper limb | 3 healthy subjects, 4 chronic stroke patients | BCI, Robotic arm, Kinect, Eye tracker/Skeleton data | SURF feature, Eye-gaze distance, 3D object maps | Robot arm aided rehab, SVM-based BCI signal classification, Lucas-Kanade object tracking |
| Chang et al. (2013) | Cerebral palsy, upper limbs | 2 patients 25 days therapy | Kinect/ SDK skeleton | Joint angles based posture | Correct exercise count through posture recognition |
| Lin et al. (2013c) | Tai-Chi upper limb rehab exercise | 2 patients with bone motor impairment | Kinect/ skeleton data | Normalised skeleton trajectory from subject and database | Grading of posture through mean error |
| Exell et al. (2013) | Stroke, arm rehab | 3 patients, 18 therapy sessions | Kinect, stimulation glove/SDK skeleton | Joint angle trajectory | Graphical comparison of Joint angle trajectory |
| Galeano et al. (2014) | Balance training system | 6 healthy subjects | Kinect, Wii balance board/ skeleton data | Mediolateral and anteroposterior sways | Rehab through FES, provide feedback through posturography |
| Su et al. (2014) | General shoulder rehab exercises | 320 vectors for training, 6 subjects for testing | Kinect/ skeleton data | Euclidean joint distance based DTW vector | Neural Fuzzy system for performance evaluation |
| Benettazzo et al. (2015) | Shoulder rehab exercises | 10 participants, 2 exercises | Kinect/ OpenNI skeleton | Joint position Euclidean Distance from reference | Audio feedback and ANN based posture recognition |
| Devanne et al. (2018a) | Low back pain rehab | 1 patient, 1 clinician, 3 exercises | 2 arm humanoid robot, Kinect/ skeleton | GPLVM based modelling space | Model clinician movement adapted to patient |
| Devanne et al. (2018b) | General rehabilitation | 5 over 60 subjects | Kinect/ skeleton | GMM, Riemannian manifolds | Classification by temporal segmentation analysis |
| Baptista et al. (2019) | Stroke, lower, upper limbs | 10 healthy subjects, two sessions | Kinect/ Depth image based skeleton (Shotton et al., 2011) | Joint angle comparison, Euclidean distance | Real time feedback |
| Schez-Sobrino et al. (2019) | Stroke, upper limbs | 1 patient, rehab exercises | Kinect/ Skeleton data | OE-DTW (Tormene et al., 2009) | Classification by time series comparison |

| | | | | | |
|---|---|---|---|---|---|
| Gu et al. (2019) | PD, whole body | 3 patient's rehab exercises | Kinect/ Skeleton data | Clinician evaluation | Graphical representation |

Table 2.3: Direct rehabilitation: Instead of virtual performance subject's physical movements are tracked for guiding or assessing rehabilitation. ANN: Artificial Neural Network, BCI: Brain-Computer Interface, DTW: Dynamic Time Warping, FES: Functional Electro-Stimulation, GPLVM: Gaussian Process Latent Variable Model, OE-DTW: Open-ended DTW, SURF: Speeded Up Robust Features, SVM: Support Vector Machines

In 'Direct rehabilitation' systems, there is usually an exercise regimen prescribed for patients whose purpose is to demonstrate their functional improvement. Patients may be guided through a web-based interface to perform tasks similar to 'Virtual rehabilitation' type applications. However, unlike 'Virtual rehabilitation', a subject's physical performance in the physical world is considered for further assessment or feedback. The discussion on 'Direct rehabilitation' methods is split into two parts: First, where CV sensor is exclusively used to obtain primary data and second, where non-vision systems such as assistive robots are used.

### 2.5.1 Pure vision-based


Figure 2.3: An instance of 'Direct rehabilitation' systems where a patient's performance is directly assessed through joint position tracking. In Lin et al. (2013c), Tai-Chi exercise pose is compared to a standard pose and feedback is provided.

Similar to 'Virtual rehabilitation' systems, early models before the advent of Kinect used indirect methods to track joints. Ghali et al. (2003) tracked hand movements by tracking objects held in hand where sequence of hand movement determined whether an activity was successfully completed. The authors used colour histograms for detecting objects based on their colour. Colour-based methods are very prone to background noise and variation in illumination (Forsyth; Ponce, 2012). Zariffa; Steeves (2011) first used background subtraction and morphological filtering to process hand grip images as first step for hand grip classification. Then, Hu invariant and contour signature extracted from the processed images were used as features for classification through K-Nearest Neighbour (KNN). Hu moments representations are not orthogonal and therefore have redundancy (Arafah; Moghli, 2016). Zariffa; Steeves (2011) have used 7 Hu moments whereas Arafah; Moghli (2016) show that 12 moments are needed to have good invariance and therefore robustness. Santilli; Laneve (2011)

show that both morphological operations and moment-based methods are more prone to noise as compared to ANNs.

The Kinect SDK provides advanced information such as kinematics and gesture recognition in addition to providing skeleton information (Han et al., 2013). Authors have used this to count the number of times correct posture was attained as a measure of rehabilitation progress (Huang, 2011; Chang et al., 2011; Chang et al., 2013). Lin et al. (2013c) used Tai-Chi exercise for rehabilitation where patients were required to attain certain Tai-Chi postures. Postures attained by a patient was compared through mean error of nine joint angles and positions with respective target postures and subsequently graded. The mean error does not tell the deviations for each joint and it is possible that deviations arise from a joint not involved in the exercise. Moreover, two different postures can have the same mean error causing the grading system to give incorrect results. To measure deviations, there are better techniques such as ANOVA, linear regression and so on (Christensen, 2018). Contemporary research on gesture/posture recognition benefit from advanced statistical algorithms such as Conditional Random Field (CRF) (Wang et al., 2006a) or DL-based algorithms (Oyedotun; Khashman, 2017). Assessment based on kinematic parameters can be more robust and accurate as compared to visual assessment by clinicians (Bigoni et al., 2016). However, in an interesting departure from the trend (Table 2.3), Su et al. (2014) proposed a Kinect-based fully independent home rehabilitation system that used clinicians experience to model a Fuzzy Logic-based neural system. The authors (Su et al., 2014) argue that clinicians grade a patient's activity as 'good', 'bad', 'slow' and so on, instead of providing a binary feedback (right or wrong) or a numerical score. Modern DL networks are capable of providing such output ('good', 'bad', 'slow') instead of binary or numerical score. It is not clear what advantage the authors gain by using a Fuzzy neural system as compared to a regular DL network.

### 2.5.2 Multimodal

In multimodal applications, CV sensors such as Kinect are combined with other assistive technologies such as assistive robots, electrical stimulation etc. Normally, patients using the rehabilitation systems are guided via visual animation or clinicians. Galeano et al. (2014) used Functional Electrostimulation (FES) for assistance while providing visual feedback through posturography on skeletal data. Frisoli et al. (2012) introduced a gaze independent, wearable Brain-Computer Interface (BCI) driven robotic exo-skeleton for upper limb rehabilitation in stroke patients. The first objective was to select real-world object by estimating eye-gaze through a CV-based eye-tracking system. The second objective was to assist patient arm movement for moving real world objects. For this, the signal from BCI was fed to a SVM classifier to ascertain if the subject intended to move his or her arm. Then, this signal was used to actuate the robotic-arm. Speeded-Up Robust Features (SURF) (Bay et al., 2006) was used for object matching and Lucas-Kanade tracking algorithm (Lucas; Kanade, 1981) was applied to track objects using depth data from Kinect. SURF is a faster alternative to Scale-Invariant Feature Transform (SIFT) whereas SIFT is more robust key-point detection algorithm (Bay et al., 2006). With the increased computation speed in modern computers, SIFT is a more appropriate key-point detection method. Since its introduction in 1981, Lucas-Kanade tracking algorithm has been widely used to calculate optical flow. Horn-Schunck (Bruhn et al., 2005) and Two-frame motion

estimation method (Farnebäck, 2003) are other widely used optical flow-based tracking algorithm that can be used in such a scenario. Another example of multi-modal system is a humanoid robot-guided system for rehabilitation from lower back pain (Devanne et al., 2018a). Devanne et al. (2018a) used a Gaussian Process Based Latent Variable Model (GP-LVM) to model clinician's movements according to the patient's morphology for guiding the rehabilitating patient. GP-LVM is a feature encoding technique that uses latent variables to reduce the dimensionality of the real variables. GP-LVM can be assumed as more generalised non-linear Principal Component Analysis (PCA) with an assumption of Gaussian Prior (Lawrence, 2004). Non-linear PCA, Auto-encoders (Baldi, 2012) are other techniques used in such situations.

Huang (2011), Chang et al. (2011), and Chang et al. (2013) had the goal of counting correct postures by calculating the joint angles. However, counting posture did not show the amount of deviation form correct posture for unsuccessful attempts. A slightly better way is to compare joint angle trajectories as in Exell et al. (2013) or grading of error through mean error as done by Lin et al. (2013b). To judge if an exercise is executed correctly, it is also essential to qualify the starting posture as correct (Benettazzo et al., 2015), which is not the case in approaches mentioned above. These approaches are mostly primitive and lack analysis of the whole temporal sequence. Later approaches have taken advantage of sequence comparison algorithms such as DTW or variants of it like OE-DTW (Schez-Sobrino et al., 2019) and combine these with various grading methods for better understanding of a patient's state. In automated rehabilitation, it is not always feasible to be guided via screen interface. In such scenarios, other assistive technologies like BCI and human motion imitating robots are very useful (Fasola; Matarić, 2013). For assistive robots, it is important to work according to the patient's morphology as demonstrated by Devanne et al. (2018a). The authors also show us very good implementation of latent model needed to transfer low-dimensional latent space to a high-dimensional robot space through GP-LVM-based probabilistic model.

## 2.6 Comparison

| Author | Target | Dataset | Raw data | Feature | Objective |
|---|---|---|---|---|---|
| Goffredo et al. (2009) | Sit to stand | 5 subjects and 5 healthy elderly | Joints marked in first RGB frame | Tracking by Gauss-Laguerre transform algorithm | Discriminatory analysis between young and old |
| Leu et al. (2011) | Gait abnormality | 20 healthy, 10 patients | 2 RGB camera system | skeleton extraction from silhouette | Knee angle based gait comparison |
| Stone; Skubic (2012) | Older adults, Gait abnormality | 5 persons, 3 weeks walks in 4 homes | Kinect, Vicon/ skeleton data | gait parameters: speed, stride time and length | Graphical analysis |
| Scherer et al. (2012) | Stroke, brain activity | 3 healthy subjects, hand closing and opening | Kinect, EEG/ NITE Skeleton data | Hand posture, EEG signals | Assessing (sub)cortical reorganisation for hand movement |
| González et al. (2012) | Balance assessment | 2 healthy subjects, 42 postures | Kinect, Wii balance board/ OpenNI skeleton | Center of position, statistically equivalent serial chain | Visual comparison of Center of Mass |
| Kurillo et al. (2013) | General upper limb reachable workspace | 10 healthy subjects, 1 FHSD patient | Kinect/ skeleton information | Joint position, angle, ROM | ANOVA analysis, statistical comparison |
| Wang et al. (2013b) | PD, upper and lower limbs | MSR-3D dataset (Li et al., 2010), 1 healthy, 1 PD subject | Kinect/ SDK skeleton | Temporal skeleton trajectory | Action recognition through TASS for extracting clinically relevant features |
| Paiement et al. (2014) | Gait, knee injury | SPHERE-Staircase2014 dataset: 48 sequences, 12 healthy, 3 knee-injury subjects | Kinect/ SDK, OpenNI skeleton data | Joint positions, velocities, pairwise distance, angles | Abnormal gait detection from statistical modelling |
| Spasojević et al. (2015) | PD, Gait, shoulder | 12 PD patients | Kinect/ skeleton data | 9 MPIs on ROM, speed, rigidity, symmetry ratio LDA | SVM, KNN, MLP based classification, Graphical Co-relation between MPIs and UPDRS |
| Leightley et al. (2015) | Young and adult standard clinical tests | 54 subjects, 13 rehab exercises | Kinect/ skeleton data | K-means clustered poses based on Manhattan distances | Pose recognition through ANN, RF, GRBM, SVM, Kinematic comparison |
| Han et al. (2015) | FHSD, upper body ROM | 22 patients, 24 healthy subjects, standard ROM tests | Kinect/ skeleton data | Normalised body-centric 3D hand trajectory | Statistical comparison |
| Tao et al. (2016) | Stroke and Parkinson's, gait, sit and stand | SPHERE datasets (Paiement et al., 2014; Tao et al., 2016) | Kinect/ SDK, OpenNI, skeleton data | Joint positions, velocities, pairwise distance, angles | HMM model with discriminative classifier based online assessment |
| Antunes et al. (2016) | Stroke, General rehab movements | ModifyAction, Weight&Balance, SPHERE-Walking2015 (Paiement et al., 2014) datasets | Kinect/ Skeleton data | Normalised and temporally aligned skeleton sequence and standardised template | Human interpretable feedback to better match standard template |
| Vakanski et al. (2016) | General motion modelling | UTD-MHAD dataset (Chen et al., 2015) | Skeleton data | DTW aligned skeleton sequence, Mean log-likelihood | MDNN to predict mean log-likelihood as performance measure |

| | | | | | |
|---|---|---|---|---|---|
| Natarajan et al. (2017) | Gait analysis | 20 healthy and 4 subjects with walking issues | RGB-D camera, Morphologically extracted skeleton from silhouette | Gait parameters, step-length, stride-length | Statistical analysis |
| Dolatabadi et al. (2017) | Stroke, upper body compensatory movement | 10 healthy, 10 stroke patients | Haptic robot, Kinect/ SDK2 skeleton data | Joint angle trajectory | Comparison based on AUC |
| Spasojević et al. (2017) | PD, Gait,shoulder and Palm | 30 PD patients, various stages | Kinect, sensor glove/ skeleton data | 25 MPIs based on Kinematic parameters, LDA | SVM, MLP and KNN classification of disease stage, Graphical Co-relation between MPIs and UPDRS |
| Baptista et al. (2017) | Physical activity assisting exercises | UTKinect dataset (Xia et al., 2012b), stroke patient balance simulation | Kinect V2/ skeleton sequence | SS-DTW and TCD based temporal sequence alignment | Visual feedback for corrective action |

Table 2.4: Comparison type applications: Articles on patient monitoring applications that provide graphical or statistical comparison of patient action but do not provide a decisive patient assessment or score. ANN: Artificial Neural Network, ANOVA: Analysis of Variance, DTW: Dynamic Time Warping, EEG: Electroencephalograph, GRBM: Gaussian Restricted Boltzmann Machines, HMM: Hidden Markov Model, KNN: K-Nearest Neighbour, MDNN: Mixture Density Neural Networks, MLP: Multi-Layer Percepteron, MPI: Movement Performance Indicator, LDA: Linear Discriminant Analysis, RF: Random Forest, ROM: Range of Motion, SS-DTW: Sub-sequence DTW, SVM: Support Vector Machines, TASS: Temporal Alignment Spatial Summarisation, TCD: Temporal Commonality Discovery

Table 2.4 summarises 'Assessment' type applications which have presented comparative analysis of kinematic data obtained from CV-based sensors. In such systems, there is no rehabilitation program designed for patients. Authors have drawn comparison ranging from simple graphical visualisation and statistical techniques to more advanced ML algorithms. This discussion is split into three parts. The first part discusses articles where kinematic data is directly used for comparison. Graphical and statistical comparison highlight differences between a patient's and a healthy subject's parameters. Secondly, applications where statistical (ML) algorithms have been used for modelling kinematic data are reviewed. The third part discusses the use of stochastic (DL) algorithms for comparative analysis of a patient's activity.

### 2.6.1 Kinematics-based modelling

In this type of applications, authors have used kinematic parameters directly for comparison. Before the introduction of Kinect, authors used other CV-based algorithms to extract the skeleton. Leu et al. (2011) used two cameras for filming 20 subjects against a standard background with the human silhouette extracted through background subtraction and image segmentation. This data was compared to a standard stick-figure model for extracting skeleton. For accuracy, the algorithm was tested against a standard sensor-based marker. Simple graphical comparison showed a visible difference between knee angle trajectories of regular and irregular gait. Natarajan et al. (2017) also used their own tracking algorithm while introducing Reha@Home. The authors argued that detection on lower extremity joint in Kinect is not accurate enough and used depth information in combination with morphological operations to extract human silhouette. Performance of the system

was evaluated through comparison with data from Electrogoniometer. Graphical trajectories of the gait parameters show visible difference between healthy subjects and patients. Leu et al. (2011) extract the human body through background subtraction from 2D image while Natarajan et al. (2017) used depth continuity from RGB-Depth image. Depth-based image segmentation can be more robust to background noise, cluttering, colours etc. than RGB-based background segmentation as depth information is not affected by background colour or varying illumination.

The TRSP dataset (Dolatabadi et al., 2017) presents 3D joint positions consisting of upper arm movements for both stroke patients and healthy subjects. Kinect was used for tracking joint positions of 10 healthy subjects and 10 stroke survivors having restricted arm movements. Two experts were recruited to annotate the dataset and the dataset was labelled into 3 different compensatory movements and one normal movement. Area Under Curve (AUC) values obtained from joint angle trajectory show substantial measurable difference between good and abnormal examples.

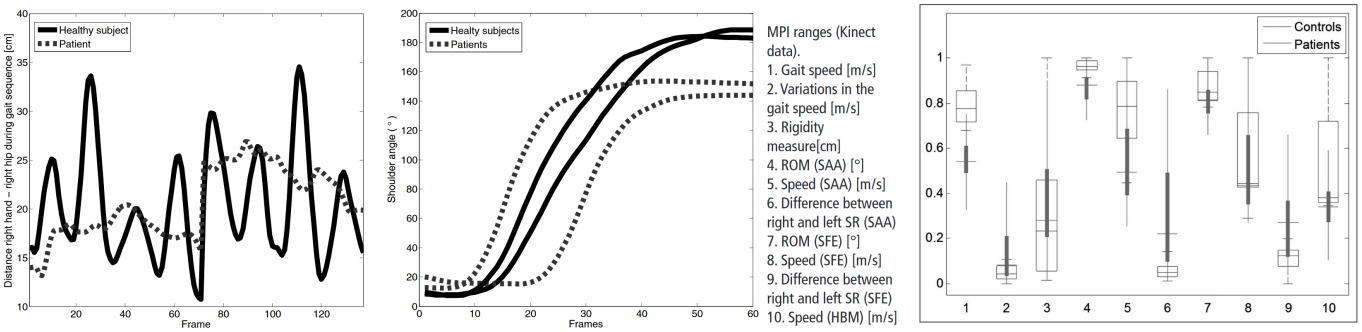



Figure 2.4: Graphical comparison of patients and healthy subjects through kinematic parameters and joint angle trajectories (Spasojević et al., 2017).

Spasojević et al. (2015) used four different body movements and measurements, for discriminating PD patients from healthy subjects. Nine 'Movement Performance Indicator' vectors were derived from different kinematic parameters related to Gait, Shoulder Abduction Adduction, Shoulder Flexion Extension and Hand Boundary Movements. Graphical and statistical comparison based on kinematic parameters showed visible differences between patients and healthy subjects. However, graphical comparisons were only made for one kinematic parameter (e.g., angular velocity of right arm) at a time. Graphical comparison is useful for simple cases involving one kinematic parameter at a time. But when assessment includes complex movements involving multiple joints a simple graphical comparison is inadequate. Thus, to assess a patient's overall condition, the nine 'Movement performance indicator' vectors were used as feature for four different classification algorithms KNN, Multi-layer Perceptron (MLP), SVM and Radial Basis Function (RBF) among which MLP and SVM performed better than the other two. It is worth mentioning that MLP is a very basic version of today's ANNs with only linear threshold activation functions (Goodfellow et al., 2016). Also, it is not clear how the authors (Spasojević et al., 2015) used RBF for classification as RBF is not a classification algorithm but a kernel that is normally used with algorithms such as SVM (Bishop, 2006).

### 2.6.2 Statistical modelling

Instead of directly comparing kinematic data, authors have also used statistical (ML) algorithms for modelling human movement, which is subsequently compared statistically or graphically. Tao et al. (2016) used HMM modelling, for online quality of motion assessment of gait on stairs, walking on flat surface, sitting and standing. The authors use a robust manifold learning technique for dimensionality reduction initially proposed by (Paiement et al., 2014). Here, the authors take original diffusion maps (Coifman; Lafon, 2006) and introduce an indicator function with weighting factor similar to Gerber et al. (2007). The indicator function helps to avoid disconnected components in Laplacian Eigenmaps that reduces the influence of outliers. The authors (Tao et al., 2016; Paiement et al., 2014) show that the manifold learning technique based on modified diffusion maps captures the intrinsic cycle nature of gait patterns better than diffusion maps. For discriminating skeleton sequences using HMM, entire sequences have to be fed to the model. This is not possible in the case of online assessment and thus, a variable window approach (Narasimhan et al., 2006) was adapted to address the problem. Four different HMM models were developed and are used to extract features from skeleton data to classify abnormalities using SVM.

Wang et al. (2013b) devised a series of exercises for musculo-skeletal patients targeting PD patients with activities include walking, walking with counting and sit to stand. The proposed Temporal Alignment Spatial Summarisation (TASS), first, segments repetitive skeletal motions from a continuous stream into Skeletal Aaction Unit (SAU)s. The SAUs are then temporally aligned against a reference SAU and spatially summarised into repetitive SAU by filtering out small variations and large outliers. TASS algorithm is inspired by Dynamic Manifold Warping (DMW) (Gong; Medioni, 2011), which is technically a spatio-temporal manifold modelling with latent variables. To calculate similarity between two skeletal sequences the authors (Wang et al., 2013b) calculate a Linear Least Squared Error combined with RANSAC (Fischler; Bolles, 1981). The method was evaluated against the standard MSR-Action3D (Li et al., 2010) action recognition dataset. However, for clinical validation, only a single PD patient and a healthy subject were asked to perform walking and sit-to-stand experiment. Validation on only a single patient subject is perhaps not enough to demonstrate the efficacy of any method in clinical situations.

Antunes et al. (2016) framed the assessment problem as feedback to be provided to a skeleton sequence to better match a standard execution sequence template. The system has been evaluated on three publicly available datasets UTKinect (Xia et al., 2012b), SPHERE-Walking2015 (Paiement et al., 2014) and Weight&Balance (Antunes et al., 2016). The authors (Antunes et al., 2016) used data normalisation for spatial alignment and DTW for temporal alignment. A rotation matrix is computed that minimises the Euclidean distance between a reference and the current sequence consisting of frame-wise skeleton information. This rotation matrix is used to provide visual feedback to the subject performing the current sequence. Antunes et al. (2016) show that by following the Euclidean distanced-based loss, the error function converges when visual guidance is correctly followed. Baptista et al. (2017) also saw the problem as essentially finding the difference between two skeleton sequences. The authors used Sub-Sequence Dynamic Time Warping (SSDTW) (Müller, 2007) and Temporal Commonality Discovery (TCD) (Chu et al., 2012) algorithms to match user action to a specific

template and provide feedback highlighting deviations from normal execution. A sequence was considered to be a match to the template if the difference between the two was lower than a preset threshold.

Baptista et al. (2017), Antunes et al. (2016), Wang et al. (2013b), and Tao et al. (2016) have all used publicly available datasets. These datasets have larger samples than articles reviewed in the previous sections (Sec. 2.4 and 2.5). Larger datasets enable the use of intelligent statistical (ML) algorithms as smaller datasets cause these algorithms to over-fit (Bishop, 2006). Also, using publicly available datsets enable future authors to evaluate their model against the current standards. Antunes et al. (2016) and Baptista et al. (2017) base their temporal alignment on the DTW algorithm while TASS (Wang et al., 2013b) is based on DMW. DMW is a manifold learning based spatial-temporal alignment algorithm whereas DTW only temporally aligns sequences. Experiments conducted by Gong; Medioni (2011) demonstrate the temporal part of DMW, and therefore the overall DMW, to be more effective than DTW. Tao et al. (2016) use HMM for sequence modelling which authors (Palma et al., 2016b) have found more effective than DTW. Euclidean loss used by Antunes et al. (2016) is more sensitive to outliers as compared to absolute difference used by Baptista et al. (2017) because squaring of error penalises larger errors more. Therefore, Baptista et al. (2017) needed to set a threshold for outlier detection manually. Both authors could have taken the advantage of the well-known RANSAC (Fischler; Bolles, 1981) for distance matching as in Wang et al. (2013b).

### 2.6.3 Introduction of stochastic methods

The area of CV-based rehabilitation and assessment has not seen an extensive application of DL. However, Leightley et al. (2015) introduced the ANN which is a simpler version of modern DL models. Leightley et al. (2015) presented the Kinect 3D Active (K3D) dataset, which captured motions based on common clinical assessments used to determine altered patients' movements. 54 subjects aged 18 to 81 were asked to perform 13 clinical tests such as balance, open and closed eyes, jump, chair stand etc. Owing to the diverse age-related conditions, a subject's movements varied widely for any given activity. Several algorithms were used for action determination out of which SVM and ANN achieved the best accuracy. To assess clinical condition the activities were further compared in terms of average time taken to complete the action. Discrimination between well-performed and poorly performed action was done on the basis of the standard deviation method proposed by Baumgartner et al. (1998). In the absence of large scale publicly available datasets, simulating or generating data has also been considered. Vakanski et al. (2016) trained their Mixture Density Neural Network (MDNN) on the standard action recognition UTD-MHAD dataset (Chen et al., 2015), to model human movement for each action. Mean log-likelihood of observed sequences was used as a performance metric in evaluating the consistency of a subject's performance. Then, random noise was imparted to generate deviations from standard action and these deviations were measured. The proposed model was programmed to be usable with Kinect captured skeleton data.

The articles presented above have more robust approach as compared to most of the 'Rehabilitation' approaches (Sec. 2.4 and 2.5) in the sense that authors have compared more kinematic parameters, used more advanced statistical analysis and have used bigger datasets. For example, in rehabili-

tation type applications, many authors have chosen a simple joint angle or joint angle trajectory comparison (Huang, 2011; Exell et al., 2013). In contrast, 'Comparison' type applications sees the implementation of more robust kinematic parameters such as 25 different 'Movement performance indicators' (Spasojević et al., 2015; Spasojević et al., 2017), gait parameters such as stride length or step length (Natarajan et al., 2017), velocity and relative position (Paiement; Tao, 2014), normalised sequences (Han et al., 2015) and so on. As a result, such applications are able to carry out more complex comparison such as gait analysis, compensatory movements and so on, as opposed to simple gesture or posture recognition based on a single joint.

One of the goals of assessment applications is to accurately represent temporal skeleton sequences for comparative analysis. One criterion for accurate representation is view-invariant representation. To this end, Natarajan et al. (2017) used human silhouette to project 3D information form depth information to 2D plane. Such processes are computationally expensive as they require a number of morphological and other image-based operations. Natarajan et al. (2017) argue that it is more robust for gait analysis than Kinect-based 3D positions which are not reliable for lower extremities. However, another option is to use HMM-based models (Paiement et al., 2014; Tao et al., 2016) which also capture view-invariant representation of the human skeleton. However, HMM training is also computationally expensive and requires a large amount of data. Computational complexity can be reduced by using latent variable modelling like GP-LVM (Devanne et al., 2018a) or by using dimensionality reduction techniques. Since the human skeleton sequence is highly non-linear, linear dimensionality reduction methods like PCA are not feasible. Instead authors have proposed non-linear techniques such as manifold learning-based on modified diffusion maps (Paiement et al., 2014; Tao et al., 2016), LSTM-based Auto Encoders (Vakanski et al., 2016), Linear Discriminant Analysis (LDA) to select more relevant MPIs (Spasojević et al., 2015; Spasojević et al., 2017). Another important aspect of skeleton sequence comparison is temporal sequence alignment, which is necessary to make a fair comparison. Many authors have temporally aligned sequences using the DTW algorithm (Antunes et al., 2016) or its variants such as SSDTW (Baptista et al., 2017). Baptista et al. (2017) also use TCD algorithm and show that the results are similar to that of SSDTW. On the other hand, Wang et al. (2013b) have used TASS based on manifold learning technique and demonstrate better performance than DTW.

The main goal of 'Comparison' type applications is to present graphical, statistical or other comparison of the subject's kinematic parameters to establish the extent of abnormality in motion. Some applications have presented simple graphical comparison of parameters such as joint angle trajectory (Leu et al., 2011; Stone; Skubic, 2012). In such situations, it is beneficial to add statistical comparison parameters such as mean error, root mean square deviation, Spearman Rank Co-Relation (Natarajan et al., 2017) or more advanced analysis such as ANOVA (Kurillo et al., 2013). This enables more robust comparison of kinematic parameters. For comparison of more complex temporal sequences involving multiple joints, simple kinematic parameter comparison is inadequate. For such situations authors have implemented more advanced approaches like log-likelihood from temporal sequence modelling (Vakanski et al., 2016; Paiement et al., 2014), clinical score co-relation with LDA reduced MPIs (Spasojević et al., 2017). Non-linear dimensionality reduction and temporal sequence modelling helps to process more complex sequences, generalise and reduce influence of outliers thereby

enabling richer and robust comparison. 'Comparison' type applications also see the introduction of publicly available datasets which paves the way for competitive evaluation of the proposed models (Paiement et al., 2014; Tao et al., 2016; Baptista et al., 2017). However, statistical comparison do not provide decisive scoring or classification of a patient's condition. The next two sections discuss applications that can classify or grade patient's physical impairment.

## 2.7 Categorisation

| Author | Target | Dataset | Raw data | Feature | Objective |
|---|---|---|---|---|---|
| Cho et al. (2009) | PD gait recognition | 7 PD, 7 healthy subjects | RGB camera | PCA and LDA from binary silhouette | Minimum distance classification |
| Taati et al. (2012) | Posture compensation | 7 healthy subjects simulating compensation | Kinect/ skeleton data | 3D orientation of subset of joint lines | Posture classification based on HMM and SVM |
| Metcalf et al. (2013) | Stroke, hand rehabilitation | 2 hand modes, 76 videos, 1692 frames | Kinect/ depth data | Contour, Kinematics-based hand model | Hand grip classification, Hand kinematics |
| Leightley et al. (2013) | General rehab activities | 20 subjects, 200 activities, 60225 frames | Kinect/ SDK skeleton data | Position, velocity, energy | Activity recognition by PCA-SVM, RF |
| Kertész (2013) | General rehab exercise, whole body | 7 subjects, 8 exercises 654 samples | Asus Xtion/ NITE skeleton data | skeleton trajectory based reference model | SVM, Numerical model based Posture recognition |
| Jun et al. (2013) | Knee Osteoarthritis | 5 healthy subjects, 2 sets * 4 trials | Kinect/ skeleton data | Normalised and PCA reduced skeleton information | Subject classification using KNN, automation for individualisation of exercise regimen |
| Liu et al. (2013) | General rehab exercise, arm | 10 subjects, 3200 postures | Kinect/ OpenNI skeleton | 45 normalised 3D coordinates | Posture recognition, Action recognition through SVM |
| Kargar et al. (2014) | Gait severity, Go and get up test | 12 elderly subjects, 50 samples | Kinect/ skeleton data | Gait features: step distance, duration; skeletal features: joint distance, angle | Fall risk classification through SVM and BoW |
| Cary et al. (2014) | Stroke, upper limbs | 10 subjects, 5 poses, 5 repetitions | Kinect/ skeleton data | 17D vector, 4 joint angles and body inclination angle | Gesture recognition through ANN |
| González-Ortega et al. (2014) | Cognitive psychomotor exercises | 10 healthy, 3 frontal lobe injury, 2 dementia, 5 tests, 14 exercises | Kinect/ OpenNI skeleton, depth image | Joint position, Mean and Gaussian curvature from depth image | Skeleton, Face eye, ears, nose detection for posture recognition using HK classification (Besl; Jain, 1985) |
| Palma et al. (2016a) | General rehabilitation exercises, upper, lower limbs | 14 subjects, 10 exercise, 100 incorrect | Kinect/ Skeleton data | Quantized joint angles | Error recognition using HMM and MD-DTW |
| Capecci et al. (2016) | General motor disabilities, whole body | 19 healthy, 14 patients, 5 Lower back pain exercises | Kinect V2/ skeleton data | Relative joint angle, velocity, constraints | HSMM for discriminating improper execution |
| Richter et al. (2017b) | Errors in hip abduction | 3 patients, 3 scenarios | Kinect/ skeleton data | Local and hierarchical coordinates based IDTW | SVM based error classification, live feedback |

| | | | | | |
|---|---|---|---|---|---|
| Leightley et al. (2017b) | Motion instability identification | K3D dataset (Leightley et al., 2015) | Kinect/ skeleton data | Joint groups comprising body centric coordinate positions, angles | Unstable motion classification using CNN and ML algorithms |
| Richter et al. (2017a) | Errors in hip abduction | 1 patient | Kinect/ skeleton data | automatically determined class specific joint combinations | Weight-based hip abduction error classification, visual feedback |
| Leightley et al. (2017a) | Standard clinical tests, whole body | K3Da Dataset (Leightley et al., 2015) | Kinect/ skeleton data | group of joint angle, distance, body lean angle | Mobility classification through ML algorithms |
| Pogrzeba et al. (2018) | Stroke, PD, repetitive hand movements | 10 healthy, 20 patient subjects, drum beats 122-126 per minutes | Kinect, SDK/ skeleton data | PCA reduced skeleton trajectories, mean, standard deviations | Logistic regression based correct speed, consistency, variability classification |
| Khan et al. (2018) | Infants with motor disabilities | 10 infants, supine and prone position video | Kinect/ RGB, depth data | size, area, position of bounding box through Body segmentation | Infant movement classification through SVM |
| Rivas et al. (2018) | Stroke patient engagement | 5 patients, 10 sessions, gesture therapy | Camera pressure gripper/ colour marker | Kinematic features, grip pressure features | MSNB classifier for patient states: anxiety, pain, engagement and tiredness |
| Chen et al. (2018b) | General upper limb rehab exercises | 20 samples, 11134 frames | Kinect/ RGB, Depth | Skeleton extraction algorithm, skin color detection for face orientation | DTW based correct exercise recognition |
| Zhi et al. (2018) | Stroke, upper body compensatory movement | TRSP dataset (Dolatabadi et al., 2017) | Kinect/ skeleton data | Noise reduced and body centric joint positions | Classification through SVM and LSTM |
| Antunes et al. (2018) | Senior lower body fitness | 11 young, 10 elderly subjects, 4 exercises | Kinect/ RGB, skeleton data | RGB video, skeleton trajectory | ANN, LSTM-based activity recognition, Kolmogorov-Smirnov test |
| Li; Vakanski (2018) | General rehabilitation exercises | UI-PRMD dataset (Vakanski et al., 2018) | Kinect/ skeleton data | Scaled and mean shifted joint angle trajectory, RMS based soft-labels | Modelling and evaluation of movement through GAN |
| Williams et al. (2019) | General rehabilitation exercises | 10 subjects, shoulder abduction, deep squat | Vicon Optical Marker/Pose | PCA, Auto Encoder-based dimensionality reduction | GMM, DTW, Distance-based classification |
| Fu et al. (2020) | Stroke | 10 Stroke patients, Compensatory movement detection exercise | 2 RGB camera/ Openpose (Cao et al., 2017) body-pose | Subset of body-pose | K-Means and DT classification |

Table 2.5: Categorisation type assessment applications: Articles that discriminate a patient's activity as correct-incorrect or provide a discrete rating. ANN: Artificial Neural Network, BoW: Bag of Words, CNN: Convolutional Neural Network, DTW: Dynamic time warping, GAN: Generative Adversarial Network, GMM: Gaussian Mixture Model, HMM: Hidden Markov Model, HSMM: Hidden Semi-Markov Model, IDTW: Incremental DTW, KNN: K-Nearest Neighbour, LSTM: Long Short-Term Memory, MD-DTW: Multiple Dimension DTW, MDC: Minimum Distance Classification, ML: Machine Learning, MSNB: Multi-Resolution Semi-Naive Bayesian, PCA: Principal Component Analysis, RF: Random forest, SVM: Support Vector Machines

The primary goal of the articles presented in this section is to categorise a patient's activity into discrete categories, such as correct/incorrect posture, good/bad movement and so on. In contrast to comparative analysis, articles reviewed in this section are more decisive in terms of providing

patient assessment. Technically, most of the articles in this section have the goal of posture or action recognition where discrimination is done between improper and proper execution of movements. Nevertheless, it also includes disease severity classifications, determination of a patient's cognitive abilities and so on. The discussion is split into two parts: 1) Statistical algorithms and 2) Stochastic algorithms-based categorisation.

### 2.7.1 Statistical algorithms-based

In 'Categorisation' type systems authors have extensively used advanced ML algorithms for categorising or classifying a patient's activity. Leightley et al. (2017a) used the K3D dataset (Leightley et al., 2015) for automated human mobility analysis where K-means clustering was used to create clinically relevant joint groups for each action. The joint groups containing relevant joint trajectories were classified using several ML algorithms for recognising the action. The authors used SVM, Random Forest (RF), ANN, Gaussian restricted boltzmann machines (GRBM), Ada-Boost, LP-Boost, RUS-Boost, Total-Boost and Bagging. Out of these, RF produced the highest average recognition rate while GRBM produced the lowest performance. Palma et al. (2016a) presented a method for detecting deviations from regular movements using HMM and Multiple Dimension DTW (MD-DTW) (Holt et al., 2007). HMM was found to be more accurate for detecting error in movements when compared to MD-DTW. Taati et al. (2012) developed an interactive system where subjects interacted with robots for posture correction where a combined HMM and SVM-based algorithm was used to recognise correct posture.

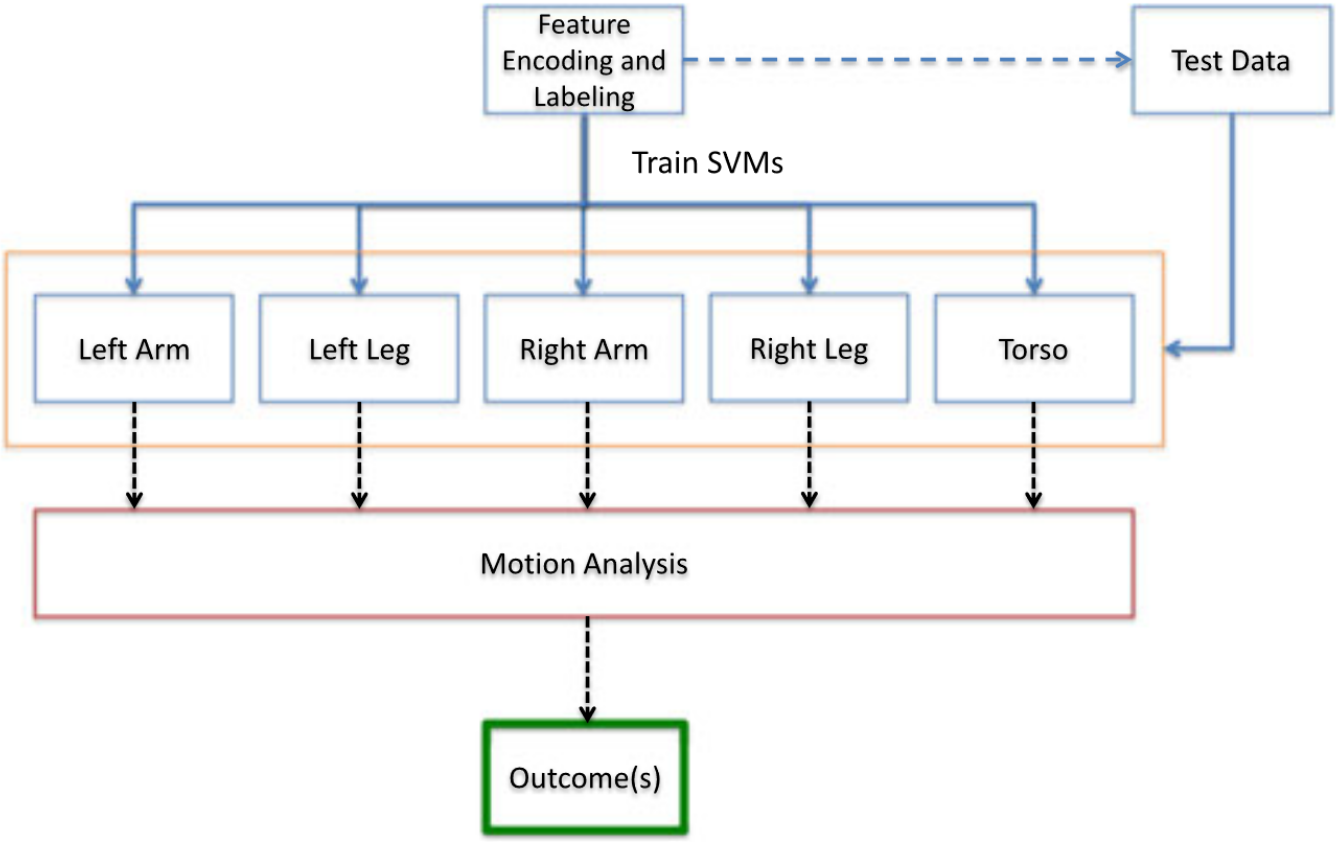



Figure 2.5: An example of Categorisation type system. Group of joints are used as encoded features for SVM. Patient's are classified as mobile or immobile. (Leightley et al., 2017a).

However, in contrast to ML algorithms, some authors (Metcalf et al., 2013; González-Ortega et al., 2014) have also used hand-crafted algorithms. Metcalf et al. (2013) used depth frames for grip classification where a hand-crafted algorithm was developed based on finger kinematics. The authors extract binary images of palm from semantic segmentation of Kinect-based depth data. The paper does not provide any details of the semantic segmentation algorithm used to extract binary images. Normally, in such situations, thresholding algorithms such as Otsu's algorithm (Otsu, 1979) may be used. Then, a contour generation algorithm extracts key-points (fingertip, finger spaces, angles) from the binary image through a geometrical kinematic model. Based on the sequence of key-points a grip classification rule was developed. Similarly, González-Ortega et al. (2014) used a set of rules based on proximity of 3D hand position to eyes, ear and nose to determine the final posture of patients for assessment of their cognitive-motor abilities. To determine the proximity, first, facial features were detected by combining pose data and depth image from Kinect with Ada-Boost algorithm (Freund et al., 1999) based on Haar-like (Viola; Jones, 2004) features. Then, eyes and nose were detected using HK classification algorithm (Besl; Jain, 1985), which is based on curvature obtained from depth image. Rule-based methods are created for very specific conditions and therefore lack generalisability and are not scale-able.

### 2.7.2 Stochastic algorithms-based

Generative Adversarial Network (GAN)s are a class of DL network that are used to generate synthetic data such as human faces (Gauthier, 2014). Li; Vakanski (2018) used the UI-PMRD dataset (Vakanski et al., 2018) to generate a synthetic dataset of incorrect human movements using GANs. Four different GAN models were trained, which included two Deep Convolutional Generative Adversarial Network (DCGAN)s (Radford et al., 2015), a Wasserstein GAN (Arjovsky et al., 2017) and a Recurrent Generative Adversarial Network (RGAN) (Esteban et al., 2017). A 1D CNN was trained as discriminator with the GANs. A soft-metric based on absolute differences was used for evaluating the performance of GANs. Li; Vakanski (2018) show a graphical comparison of the synthetic data based on LSTM. LSTM is a well-known ANN algorithm widely used for processing of sequential data. Zhi et al. (2018) used LSTM and SVM for automatic detection of compensatory movements during robotic stroke rehabilitation therapy. The authors report that both LSTM and SVM did not perform well to detect compensatory movements and cite the small size of the dataset and difficulty in maintaining exercise protocol as possible reasons. This research exemplifies the need for large-scale datasets for AI-based algorithms to achieve the level of success that these algorithms enjoy in other areas. Moreover, it also shows the difficulty in obtaining data from patients. Antunes et al. (2018) introduce the AHA-3D sequences of 3D skeletal data involving standard fitness tests on young and elderly subjects, for automatic fitness exercises assessment. The authors use LSTM-based model with joints or combination of joints and their velocities. Experiments demonstrate that the combination of joints and velocities perform better for correct exercise classification.

To summarise, as primary data, authors have mostly used only skeleton data, with few exceptions such as González-Ortega et al. (2014), Metcalf et al. (2013), and Khan et al. (2018), who have used depth data. Some authors have relied on using kinematic parameters directly as extracted features (Kertész, 2013; Kargar et al., 2014; Palma et al., 2016a; Leightley et al., 2017b). However, others

have benefited from statistical techniques for feature extraction. González-Ortega et al. (2014) used Haar features (Viola; Jones, 2004) with Ada-Boost (Freund et al., 1999) classifier. Haar features have been widely used for facial key-point detection, but recently CNN-based methods have proved to be more robust (Zhang et al., 2016; Sun et al., 2018; Yang et al., 2015b). Other examples of using statistical techniques for feature extraction include dimensionality reduction (Cho et al., 2009; Jun et al., 2013; Leightley et al., 2013), noise reduction (Zhi et al., 2018) and normalisation (Jun et al., 2013; Liu et al., 2013). Jun et al. (2013) and Leightley et al. (2013) used PCA for dimensionality reduction whereas Cho et al. (2009) used LDA for the same. As discussed in the previous section, PCA is a linear dimensionality reduction technique and may not be the best option for highly non-linear human skeleton trajectories. In such cases, non-linear techniques such as non-linear PCA, diffusion maps (Coifman; Lafon, 2006) etc. may be more suitable. Zhi et al. (2018) use Savitzky–Golay (Schafer, 2011) filter for smoothing or noise reduction. However, nowadays, noise reduction or dimensionality reduction is carried out as part of feature representation through various data modelling techniques including, but not limited to HMM and SVM (Bishop, 2006). When using such algorithms often regularisation technique is employed for outlier detection (Bishop, 2006) that helps in mitigating the effects of noise. Thus, more recently authors have extensively used advanced statistical (ML) algorithms where it is not necessary to explicitly used dimensionality or noise reduction techniques.

The task of categorisation can be split into two parts: i) modelling the data and ii) classifying the data based on the modelled representations. In this regard, authors have approached the task of categorisation in three ways. First, where there is no data modelling involved and classification is done based on kinematic parameters directly with statistical methods such as SVM, RF and so on. Second, where time sequence modelling algorithms (HMM, DTW, Multiple Dimensional Dynamic Time Warping (MDDTW), Hidden Semi Markov Model (HSMM)) are directly used for classification through mechanisms such as log-likelihood loss (for HMM) or Euclidean distance loss (for DTW) (Palma et al., 2016b; Capecci et al., 2016). In this case, algorithms exclusively used for classification have not been used. In the third case, authors have chosen to use classifiers such as SVM, RF, ANN, KNN to classify the modelled representations. Taati et al. (2012) model the data using HMM and use SVM to classify the modelled data and show that the combination of HMM and SVM works better than using SVM alone. In 'Categorisation' type applications authors have extensively used intelligent algorithms for classification, presenting an opportunity to compare these algorithms. Leightley et al. (2017a) compare SVM, RF, ANN, GRBM, Boosting and Bagging and show that RF outperforms the other methods. However, SVM is a very efficient and high performing classifier which has been widely used in this domain (Taati et al., 2012; Kertész, 2013; Leightley et al., 2013; Richter et al., 2017b) as well as in other applications of ML. Authors have also used ensemble learning techniques such as Boosting (González-Ortega et al., 2014; Leightley et al., 2017a), RF and Bagging (Leightley et al., 2017a) in 'Categorisation' type applications. Ada-Boost used by González-Ortega et al. (2014) and Leightley et al. (2017a), LP-Boost, RUS-Boost, Total-Boost used by Leightley et al. (2017a), together with XG-Boost and Gradient Boosting Machines (GBM) are commonly used boosting algorithms. Boosting, Bagging and RF are commonly used ensemble learning techniques (Bishop, 2006), which work by combining a set of weak classifiers to create a robust model which is less prone to over-fitting (Bishop, 2006). While González-Ortega et al. (2014) do not indicate why Ada-boost was chosen over other techniques, Leightley et al. (2017a) compared various LP-Boost , RUS-Boost, Total-Boost and

RF to show that RF performs better than other ensemble learning methods. Outside this domain, Rahman et al. (2020) compare boosting classifiers to recognise ADL and show that GBM achieves the best performance with full feature set. However, Ada-Boost achieves the best result with reduced feature set and thus may be more suitable for real-time performance. Hamza; Larocque (2005) and Banfield et al. (2006) compare the performance of various ensemble techniques based on Decision Trees (Bishop, 2006) on multiple datasets. While Hamza; Larocque (2005) show that RF performs better than other methods, Banfield et al. (2006) conclude that different algorithms perform better in different situations.

Unlike statistical approaches, DL approaches such as LSTM, CNN are able to both model the data and classify them efficiently. Zhi et al. (2018) show that LSTM works better as compared to SVM while Leightley et al. (2017b) show that CNN outperforms other statistical (ML) classifiers. Many authors (Cary et al., 2014; Leightley et al., 2017a; Antunes et al., 2018) have used ANN for classification, but these FC dense networks are computationally expensive and prone to over-fitting (Goodfellow et al., 2016). Using other types of DL-based networks such as CNN, LSTM are more suited and widely used for such tasks (Goodfellow et al., 2016).

## 2.8 Scoring

| Author | Target | Dataset | Sensor | Feature | Objective |
|---|---|---|---|---|---|
| Venugopalan et al. (2013) | Brain injury, upper limbs | 16 videos, 6 different poses, 4 subjects | Kinect/ skeleton data | normalised position, velocity, acceleration | Template matching and DTW for calculating similarity score |
| Hsiao et al. (2013) | Stroke, box and block test, hand | 6 trials, 100 boxes moved | Kinect/ OpenNI depth | Contour, largest circle in contour | Hand and box detection, calculate number of box transferred |
| Cuellar et al. (2014) | General rehab exercises, lower and upper body | 10 healthy subjects, 3 motion and 2 posture holding exercise | Kinect/ skeleton data | Transformed 3D vector from quaternion rotation of joint angles | Performance scoring, Template based posture matching, DTW based skeleton sequence matching |
| Khan et al. (2014) | PD, rapid finger tapping | 13 Parkinson's patient, 387 clips | RGB camera/colour marker | Motion analysis of fingers by motion-template gradient algorithm | Severity classification through SVM based on UPDRS attributes |
| Olesh et al. (2014) | Stroke hemiparesis, upper limbs | 9 patients, 10 arm movements, 5 to 28 repetitions | Kinect/ skeleton data | Joint angle trajectory | Automated scoring based on averaged temporal profiles |
| Wang et al. (2014b) | Stroke, upper limbs | 24 stroke patients, Shoudler and elbow movements | Kinect V2/ skeleton data | 9 kinematic parameters composed of velocity, energy and angle | Automated clinical scoring (FMA) through SVR |
| Dyshel et al. (2015) | LID, upper limb | 9 PD with LID patients 24 recording following AIMS protocol | Kinect/ SDK skeleton data | Single Number deducted from most discriminatory joint motion segment | Automated AIMS grading through soft SVM based algorithm |
| Ciabattoni et al. (2016b) | General rehabilitation exercises | 6 subjects, 5 exercises | RGB-D camera | quaternion based pose distance, Virtual joint angles | DTW algorithm based assessment scores |

| | | | | | |
|---|---|---|---|---|---|
| Ciabattoni et al. (2016a) | Lower back pain exercises | 5 subjects, 5 exercises | RGB-D camera/ skeleton data | joint angle, distance, torso surface | Exercise performance scoring system based on fuzzy logic |
| Soran et al. (2016) | Infants, SMA | 15 SMA patients, 72 minute film | Kinect/ Colour marker | limb trajectory | CNN-based clinical disease progression sore |
| Capecci et al. (2018) | General rehab exercises, whole body | 22 healthy, 19 neurological disability, 5 exercises | Kinect/ skeleton data | joint angles, velocity, constraint angles | HSMM and DTW model for automated clinical scoring |
| Li et al. (2018b) | Idiopathic PD, whole body | 9 PD patients, 134 footage for 4 tasks | RGB camera/ CPM skeleton data | Subset of joint positions | Classification of PD type, Regression for clinical assessment score |
| Eichler et al. (2018) | Stroke, upper limb | 12 stroke, 10 healthy, FMA mmovements | 2 Kinect V1/ skeleton data | FMA related kinematic parameters | Automated clinical scoring (FMA) through SVM and RF |
| Einarsson et al. (2018) | PD, whole body | 33 healthy, 30 patient's, | Kinect/ Skeleton data | Normalised, scaled joint coordinates | Automated UPDRS scoring based on Spare Ordinal classification |
| Capecci et al. (2019) | Stroke, PD, back-pain rehab exercise | 44 healthy 34 patients | Kinect / skeleton data | joint angles | rule, template based algorithm, Spearman Correlation |
| Liao et al. (2019) | General rehabilitation exercises | UI-PRMD dataset (Vakanski et al., 2018) | Kinect/ Skeleton data | Auto-encoder reduced skeleton sequence, GMM log-likelihood based performance metric | Automated assessment based on CNN-LSTM regressed scoring |
| Hagihara et al. (2020) | Preschool Posture control | 14 girls 3-6 year | Openpose (Cao et al., 2018) | Static Postural Balance, Antigravity Posture | Spearman Rank Correlation with clinician score |

Table 2.6: Scoring type assessment system: Articles that provide a clinical or author proposed scoring of a patient's activity. CNN: Convolutional Neural Network, DTW: Dynamic time warping, FMA: Fugl-Meyer Assessment, HSMM: Hidden Semi-Markov model, GMM: Gaussian Mixture Model, LMC: Leap Motion Controller, LSTM: Long Short-Term Memory, PD: Parkinson's disease, RF: Random Forest, UPDRS: Unified PD Rating Scale, SVM: Support Vector Machines, SVR: Support Vector Regression

In this section, articles which aim to provide an automated assessment of a patient's state are reviewed. This includes both clinical (e.g., FMA, UPDRS) and author proposed (non-clinical) scoring. For musculo-skeletal diseases, there are often a multitude of factors that describe a patient's state or condition. Simple movements such as hip abduction or individual exercises may be classified into correct or incorrect. However, to describe a patient's state, clinicians often use standard scoring systems such as FMA (Gladstone et al., 2002), UPDRS (Rating Scales for Parkinson's Disease, 2003) etc. In 'Scoring' type application the score may be either discrete or continuous which in ML terms correspond to use of supervised classification and regression respectively. This discussion is split into two parts: 1) author proposed and 2) clinical scoring.

### 2.8.1 Author proposed scoring

PReSenS, developed by Cuellar et al. (2014) is an exercise system where physiotherapists can remotely upload exercise templates to be followed by patients at home. Posture is compared to single exercise template whereas motion is compared by the time series matching algorithm DTW. Features such as joint angle, joint rotation etc. are used with DTW for action comparison. All motion (action) sequences were summarised using Piece-Wise Aggregation Approximation (PAA) for scoring performance. Keogh et al. (2001) compare PAA, with other time-series dimensionality reduction techniques such as DTW, Spectral Decomposition, Wavelet Decomposition and show that PAA performs better than other algorithms. Khan et al. (2014) used rapid finger tapping test for clinical evaluation of PD patients where subjects were asked to tap their index-finger besides their face and above shoulders. Here, the goal was to measure index-finger movements to grade patients. For automated assessment, first, a region of interest was selected as rectangle besides face. Then, face detection was achieved by Haar Cascade classifier (Viola; Jones, 2004). Index-finger was segmented through a motion-template gradient algorithm (Bradski; Davis, 2002) which contained five steps: 1) silhouette detection, 2) Motion History Image (MHI) updates, 3) motion gradient calculation, 4) motion orientation calculation and 5) motion segmentation. Kinematic parameters for calculating UPDRS features were extracted from the index-finger motion and were used classification using SVM. Normally silhouette-based methods are prone to noise but the use of MHI makes the algorithm less sensitive to silhouette noise (Tsai et al., 2015). However, with similar actions MHI may generate indistinguishable patterns and to mitigate this Tsai et al. (2015) combine the widely used optical-flow with MHI. Recently, Regional Proposal Network CNN (R-CNN)-based approaches have outperformed other approaches for image segmentation (Hariharan et al., 2014; He et al., 2017).

Venugopalan et al. (2013) proposed a real-time traumatic brain injury assessment system with two Kinect cameras and a near infra-red motion sensor used to film patients at home. Real-time patient data from the system was compared with data from observation in a clinical setting to compute similarity score. The authors used a template sequence for scoring (comparing) the patient-sequence through: i) Cross-Correlation ii) A direct frame algorithm and iii) DTW, and show that DTW outperforms the other two approaches. Liao et al. (2019) proposed a log-likelihood based performance metric to train their DL framework for assessment of rehabilitation exercises. Low-level skeleton data was represented through a deep Auto Encoder (AE) network to initially train a Gaussian Mixture Model (GMM) for calculating log-likelihood. Using the UI-PRMD dataset (Vakanski et al., 2018), the authors then trained and compared the performances of CNN, Recurrent Neural Network (RNN) and Hierarchical Neural Network (Du et al., 2015). For supervision GMM log-likelihood based performance metric was used as label to regress the network for predicting deviations from normal actions. The authors (Venugopalan et al., 2013) show that performance metric based on GMM log-likelihood works better than metrics based on Euclidean distance, Mahalanobis distance and DTW. Venugopalan et al. (2013) and Liao et al. (2019) show that model-based approaches such as DTW, GMM log-likelihood works better than non-model approaches such as Euclidean distance, Mahalanobis distance, Cross-Correlation. This is in-line with the analysis presented in the last section, which showed that modelling the data is beneficial.

### 2.8.2 Clinical scoring

The goal of assessing the state of a patient in terms of clinical scoring is a challenging task considering the multitude of factors involved in the assessment. Eichler et al. (2018) propose a two Kinect system for automated FMA, where the data from the two cameras was synchronised/aligned through body-pose stream and 3D point cloud. Kinematic parameters relevant to FMA were obtained from body-pose data and used for scoring with SVM and RF based on Decision Tree. RF based on Decision Tree performed better than SVM and authors reason that Decision Trees have in-built feature selection and thus perform better. However, the analysis in the last section 2.7 shows that SVM performs better when the data is modelled through algorithms such as HMM, DTW and so on. In Eichler et al. (2018), selecting data from one of the two cameras work better than averaged data from both the cameras. This shows that the result depends upon the orientation of the subject and defeats the purpose of having two cameras.

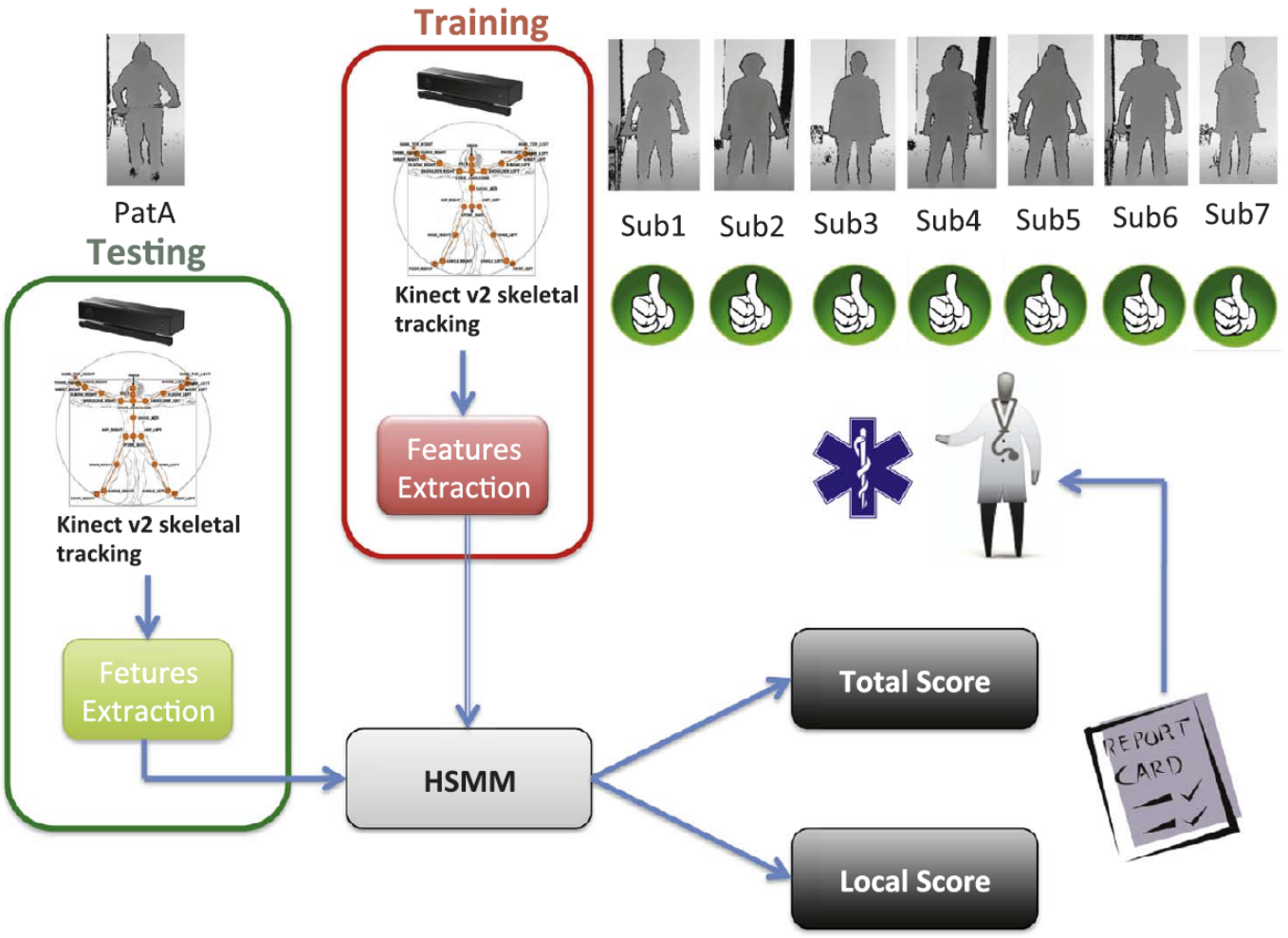



Figure 2.6: An illustration of scoring type systems. Extracted features from patient are compared to a pre-trained HSMM for automated clinical scoring (Capecci et al., 2018).

Dyshel et al. (2015) recorded Kinect-based body-pose of nine PD patients with varying severity of Levodopa Induced Dyskinesia (LID) for automated scoring on Abnormal Involuntary Movement Scale (AIMS). The authors performed motion segmentation by determining the angles between vector differences of adjacent points in motion sequences. For feature selection, chunks were extracted for each joint motion and put into distributions represented by two 30-bin histograms. One histogram

represented the normal and others represented the dyskinetic state. Earth Mover Distance (EMD) was calculated and 10 motion chunks representing the highest discrimination were selected. Each 10-dimensional vector was then reduced to a single number using one of the three methods: average motion length, average motion speed, distribution of quantised motion lengths. Soft-margin SVM-based algorithm was used to calculate AIMS score. The authors (Dyshel et al., 2015) use their own hand-crafted or rule-based approach for motion segmentation, which is very specific to the scenario involved, lack generalisation and may not be scalable. Motion history gradients (Bradski; Davis, 2002), normalised-cuts (Shi; Malik, 1998), or TCN-based approaches (Farha; Gall, 2019; Lea et al., 2017) are some of the commonly used algorithms that can be used for better generalisation. It is also not clear why EMD was preferred over other distances such as Manhattan distance, Euclidean distance and so on.

Since the introduction of DeepPose (Toshev; Szegedy, 2014) in 2014, CNN-based human pose estimation has achieved very high accuracy. Li et al. (2018b) used the 3D version of well-known CNN-based model called Convolutional Pose Machines (CPM) (Wei et al., 2016) for extracting skeleton data to analyse LID. The study involved creating a publicly available dataset involving 9 participants having LID where the skeleton data extracted using CPM was used to generate 15 kinematic features. The extracted features were normalised and then smoothed using Savitzky-Golay filter. Spectral features were computed from the Welch power spectral density (Welch, 1967) of the displacement and velocity signals. These spectral features were used to calculate the clinical rating of PD or LID severity through a RF-based regressor. The authors (Li et al., 2018b) do not provide the reason for using spectral features however, spectral features are widely used for signal processing with mechanisms such as Fourier transformation and so on. Recently, the AI community has integrated spectral features in DL models (Shaham et al., 2018; Kipf; Welling, 2016; Bruna et al., 2013) and spectral feature-based Graph Neural Networks (GNN) are emerging as an alternative to the highly successful CNNs.

In the above-mentioned articles, authors have relied mostly on skeleton data obtained from Kinect, with the exception of Li et al. (2018b), who have used CNN for pose extraction from RGB data. Most authors have used simple kinematic parameters directly as primary features. Cuellar et al. (2014) and Ciabattoni et al. (2016b) used quaternion-based pose distance as primary features. Quaternions can help in capturing rotation in 3D which is better than the normally used Euclidean distance. Liao et al. (2019) show a direct comparison between Euclidean distance, DTW distance, Mahalanobis distance and GMM log-likelihood distance as performance scoring metric and show that model-based approaches work better than model-less approaches. The authors show that dimensionality reduced sequences provide better standard deviation between correct and incorrect sequences through GMM log-likelihood metric. In most applications there is little application of dimensionality reduction techniques applied to kinematic data. However, Liao et al. (2019) show that dimensionality reduction techniques improve the performance of their application. Non-linear dimensionality reduction techniques help to reduce the computational complexity and increase effective learning and the authors show that AE outperforms PCA which is a linear-method. Dyshel et al. (2015) use PCA to improve performance but non-linear methods such as AE may have provided even better results. Like 'Categorisation' type applications, authors have used HMM or DTW for modelling the tempo-

ral data and scoring a patient's activity (Capecci et al., 2018; Venugopalan et al., 2013; Ciabattoni et al., 2016b). As explained in section 2.7 HMM and DTW are good for time-series modelling but it helps to include a classifier (SVM) or regressor (Support Vector Regressor (SVR)) to calculate scores. SVM has been most widely used to calculate but Eichler et al. (2018) show that RF based on Decision Trees outperformed SVM in their experiments. Together with results from the last section, it can be said SVM is the most popular technique but other statistical algorithms such as RF, or Bagging and Boosting etc. may provide better performance. Soran et al. (2016) have used CNN for regressing score based on images but the architecture which resembles a four convolutional layer AlexNet (Krizhevsky et al., 2012) is very basic and many recent CNN architectures can provide much better performance (Aloysius; Geetha, 2017). Liao et al. (2019) have used a combination of 1D CNN and LSTM for regressing deviation scores and show that the combined approach works better than using CNN or LSTM alone. Only two applications (Soran et al., 2016; Liao et al., 2019) have used DL methods and it remains to be seen how modern DL algorithms would perform classification when large-scale datasets are unavailable.

## 2.9 Datasets

| Author | Impairment | Details | Sensor/Data |
|---|---|---|---|
| SPHERE-Staircase2014 (Paiement et al., 2014) | Walking-up stairs | 48 sequences, 12 subjects, normal and abnormal gait | Kinect/ Open NI skeleton |
| SPHERE-Walking2015 (Tao et al., 2016) | Walking | 40 sequences, 10 subjects, normal and abnormal gait | Kinect/ Kinect SDK, OpenNI SDK skeleton |
| SPHERE-SitStand2015 (Tao et al., 2016) | Sit to stand | 109 sequences, 10 individuals, restricted knee, hip, freezing | Kinect/ Kinect SDK, OpenNI SDK skeleton |
| TRSP (Dolatabadi et al., 2017) | Stroke, compensatory movement | 10 healthy, 10 stroke 4 compensatory movements | Kinect, Haptic robot/ Kinect SDK skeleton |
| Parkinson's pose estimation (Li et al., 2018b) | PD, LID, UPDRS assessment tasks | 526 sequence, PD, LID patients, 4 UPDRS assessment tasks | RGB Camera/ CPM (Wei et al., 2016) skeleton |
| UI-PRMD (Vakanski et al., 2018) | General rehabilitation exercises | 10 subjects, 10 exercises, 10 repetitions | Kinect Vicon/ Kinect SDK skeleton |
| KIMORE Dataset (Capecci et al., 2019) | Stroke, PD, back pain exercises | 44 healthy, 34 patient subjects, 5 exercises 5 repetitions | Kinect/ RGB, depth, skeleton |
| AHA-3D Dataset (Antunes et al., 2018) | Senior lower body fitness | 11 young, 10 elderly subjects, 4 exercises | Kinect/ RGB, depth, skeleton |
| SU-10 Dataset (Negin et al., 2013) | Therapeutic gestures for balance, flexibility, strength | 12 subjects, 9 males, 1200 gestures | Kinect/ RGB, depth, skeleton |

Table 2.7: Publicly available datasets that include physically impaired activity. PD: Parkinson's Disease, LID: Levodopa Induced Dyskinesia, UPDRS: Unified Parkinson's Disease Rating Scale, CPM: Convolutional Pose Machines

Table 2.7 summarises publicly available datasets that have been captured through CV-based sensors. SPHERE (Paiement et al., 2014; Tao et al., 2016) is a series of datasets that present normal and abnormal movements for walking, walking-up stairs and sit to stand movements. Vakanski et al. (2018) introduced the UI-PRMD datasets consisting of 10 different physical activities commonly performed in rehabilitation or physical therapy scenarios. The dataset provides skeleton data obtained through Kinect along with joint angles. Mean Square Error (MSE) on joint angles has been used by authors to calculate variability between each subject, which has also provided a benchmark for establishing incorrect movements. The datasets in Table 2.7 mostly explore body-part or impairment-specific exercises which comprise of repetitive single joint movement such as elbow movements (Li et al., 2018b) or involve a few joints such as in gait assessment (Paiement et al., 2014). In contrast, ADL present more complex sequences involving several body parts which are neither repetitive nor-specific. Authors (Table 2.7) have not explored functional assessment of physically impaired persons through ADL which is widely used (Green; Young, 2001). The current research aims to improve functional assessment of patients through ADL and lack of datasets involving physically impaired versions of ADL provides the motivation for the dataset presented in Chapter 5.

## 2.10 Analysis

In this section, the algorithms and techniques used in articles reviewed are analysed in terms of their usage and drawbacks. The discussion also suggests alternatives that are more recent and may provide better performance. Research in this domain is very different from objectives like activity recognition where the common goal is to explore machine learning and pattern recognition techniques to recognise various activities. In case of areas like activity recognition, often the datasets used to evaluate the models are the same and thus a direct comparison between various methods employed by authors is useful (Ke et al., 2013; Vrigkas et al., 2015). However, due to the widely varying goals, datasets used and types of physical impairments, such comparison in this domain is difficult. Instead, this study compares the general methods and algorithms employed by researchers to achieve their goals. Following section 2.2, the discussion is split into data, feature encoding and feature comparison.

### 2.10.1 Physical Impairment Data

In articles discussed in this review, authors have mostly used Kinect-based skeleton data. The main advantage is that Kinect provides RGB videos, depth videos and 3D joint positions as well as posture through a very cheap and easy to use hardware/software system. So, authors from domains other than CV can take advantage of it. However, the Kinect system is not very accurate (Webster; Celik, 2014) and today's DL-based solution outperforms the Kinect system both in-terms of 2D (Cao et al., 2016a; Fang et al., 2017) and 3D pose estimation (Yang et al., 2018; Pavlakos et al., 2018; Pavllo et al., 2019). Due to the lack of direct comparison, it's difficult to gauge the scope of improvement in the researches above with DL-based methods instead of Kinect. Unlike other areas of CV application such as activity recognition, authors have not used RGB or depth data in combination with skeleton

information. RGB data lacks the precise joint positions whereas skeleton data lacks information such as optical flow, curves, edges etc. Modern neural networks are very good at learning such information and combining skeleton data with RGB information guides the DL model to focus on visual features on the human body. This has led to increased accuracy of activity recognition models (Baradel et al., 2018a; Tran et al., 2015). and research in this domain can also benefit from the same. Authors have also used colour based tracking, such as tracking hand while holding a coloured ball (Sucar et al., 2008b), skin colour tracking (Chen et al., 2018b) and so on. These methods were in use before the introduction of Kinect, but some of them are still in use today. They have several limitations such as tracking only one part of body and are subject to background interference etc. It also needs to be noted that Kinect is no longer in production and researchers will need to switch to other devices such the Orbec Astra (Coroiu; Coroiu, 2018). Coroiu; Coroiu (2018) discuss the interchangeability and accuracy of Orbec Astra and the Kinect device. So, it is worth taking the time and effort to switch to new devices and techniques. Authors have also used non-vision devices including, but not limited to BCI, LMC and assistive robots (Frisoli et al., 2012). A combination of vision and non-vision devices have the potential to expand the domain of CV-based physical rehabilitation and assessment.

### 2.10.2 Feature Encoding

| Method | Usage | Drawbacks | Alternatives |
|---|---|---|---|
| Colour trajectory | Track body part through coloured object | Limited skeleton tracking, prone to background interference | Skeleton tracking |
| Skeleton trajectory | Tracking body parts | Do not quantify physical characteristics | Kinematic parameters |
| Kinematic parameters | Indicates physical ability | Very specific to type of impairment(s) | None |
| Contour signature | Mark hand boundaries for grip classification | Cannot handle noisy, blurry images | DL-based segmentation (Badrinarayanan et al., 2017; He et al., 2017) |
| Hu invariant | Image boundary descriptor for grip classification | Cannot handle noisy, blurry images | DL-based segmentation (Badrinarayanan et al., 2017; He et al., 2017) |
| AUC | For comparing kinematic trajectories | AUC can be same for different curves | Statistical analyses, KLD |
| Log likelihood | Probabilistic encoding of skeleton sequence | Specific formula needed for calculate likelihood, non trivial estimation | KLD, Cross Entropy |
| SURF | Encodes local RGB features | Less accurate than SIFT although faster, cluttered key-points | SIFT, ORB |
| Depth maps | Body part segmentation, skeleton detection | Missing colour, texture, skeleton information | Use with RGB and skeleton data |
| GPLVM | Dimensionality reduction of skeleton sequence | Assumes independent distributions, needs strong prior | Non-Linear PCA, LDA, Autoencoders (*Non Linear Dimensionality Reduction*, 2020) |
| Gaussian mixture model | Encoding skeleton sequence for performance metric | Expensive for high dimensional data, need to set number of clusters | Spectral clustering, Manifold learning |

| | | | |
|---|---|---|---|
| Gauss Laguerre transform | Encoding video sequence in GLT domain | Needs manual marking to select area for transform | SIFT, SURF, ORB etc. as Key-point descriptors |
| Human body silhouette | Human body segmentation | Cannot handle noisy, blurry images | DL-based semantic segmentation (Badrinarayanan et al., 2017; He et al., 2017) |
| Pairwise skeleton trajectory | Enables relative trajectory encoding | Over-fitting, cannot learn the general trend | Learn-able encoding methods |
| K-means clustering | Encoding kinematic parameters | No of clusters needs to be manually set | GMM |
| Distances (Manhattan, Euclidean) | Encoding patient sequence distance w.r.t standard template | over-fitting, cannot learn the general trend | lear-nable encoding methods |
| PCA reduced sequence | Dimensionality reduction of skeleton sequence | Mean and co-variance does not always describe distribution | LDA, autoencoder, (*Non Linear Dimensionality Reduction*, 2020) |
| Colour segmentation | Track body part through coloured object | Prone to background noise, interference | DL-based semantic segmentation (Badrinarayanan et al., 2017; He et al., 2017) |
| GAN generated sequences | Generation of artificial data | hard to train and converge | Different types of GANs (Im et al., 2018) |
| Quaternion sequences | Represent orientation and rotation of skeleton sequence in 3D | Contains only rotation but no scaling and translation | Affine transformation matrices |
| Motion template gradient | Human motion encoding, through successive frame silhouette | Pixel-based approaches prone to background noise | Optical flow based approaches, Graph-cut algorithm |
| Autoencoder | Dimensionality reduction of sequence | Requires more data, not generally used for dimensionality reduction | Non-Linear PCA, LDA, Diffusion Maps (*Non Linear Dimensionality Reduction*, 2020) |
| LDA | Dimensionality reduction of skeleton sequence | Needs labelled data, lot of tune-able parameters | Non-Linear PCA, Auto-encoders, Diffusion Maps (*Non Linear Dimensionality Reduction*, 2020) |
| Diffusion Maps | Non-linear dimensionality reduction of skeleton sequence | Requires spectral decomposition of kernel matrix, unfeasible for large datasets | Non-Linear PCA, Auto-encoders (*Non Linear Dimensionality Reduction*, 2020) |

Table 2.8: A summary of feature encoding methods used, their drawbacks and alternatives that can be used. LDA: Linear Discriminant Analysis, ORB: Oriented FAST and rotated BRIEF, SIFT: Scale Invariant Feature Transform

Table 2.8 highlights the various feature encoding methods used by authors. It also outlines their drawbacks and suggests alternatives. Many authors have used kinematic parameters or statistics (trajectory, mean value, range of motion etc.) derived from these parameters directly as features for comparison (Kurillo et al., 2013; Stone; Skubic, 2012). Instead of directly encoding kinematic parameters, the relationship between parameters such as performance indicators (Spasojević et al., 2015), pairwise relations (Paiement et al., 2014) etc., have also been used. While such parameters are useful for simple purposes such as posture recognition, joint mobility etc., these are highly specific to the physical impairment, thus not generalisable and may suffer from over-fitting. This is because deducing the statistics form parameters does not involve modelling the data. Model-based approaches 'learn' from the data and are more generalisable and less prone to over-fitting (Bishop, 2006). Thus, a better alternative would be to learn from the data instead of comparing

kinematic parameters numerically or graphically. More recently, authors have used techniques such as DTW (Vakanski et al., 2016), HMM (Tao et al., 2016), TASS (Wang et al., 2013b) and so on to build temporal models that can help to discriminate differences between patient and ideal pose sequences. In addition to pose-based methods, authors have also used RGB videos to encode features for achieving their goal. The feature encoding techniques include Hu moments (Zariffa; Steeves, 2011), colour-based segmentation (Leu et al., 2011), motion-template gradients (Khan et al., 2014) and silhouette extraction (Cho et al., 2009). These are mostly pixel-based techniques which suffer from noise interference and do not work in case of blurry images. Modern alternatives include use of generalised local feature descriptors such as SIFT, SURF, Oriented FAST and Rotated BRIEF (ORB) or image descriptors such as Bag of Words (BoW), Histogram of Oriented Gradients (HoG) etc. Modern techniques also involve DL-based algorithms for semantic segmentation (Badrinarayanan et al., 2017; He et al., 2017) which have produced state-of-the-art results. But these require large-scale datasets which should also be publicly available for performance comparison. In the absence of large-scale datasets, using GANs for modelling artificial patient data may be very useful as shown by Li; Vakanski (2018). Although there are many variants of GANs, each of which has their own domain of applicability and limitations, Data Augmentation GAN (Antoniou et al., 2017) has been purpose-built for augmenting data. Im et al. (2018) present a quantitative comparison of various GAN types to illustrate their relative abilities. To compensate for the lack of datasets, one can also look at learning from single images (Wu et al., 2016b).

### 2.10.3 Feature comparison

| Methods | Usage | Description | Benefits | Drawbacks |
|---|---|---|---|---|
| Numerical comparison | Numerically comparing kinematic parameters, skeleton position | Joint movement, posture comparison | Lacks generalisation, statistical significance | Statistical comparison |
| Graphical comparison | Graphically comparing kinematic parameters, skeleton sequence | Comparison of normal vs patient trajectory | Lacks generalisation, statistical significance | Statistical analysis (Aanova, Chi-Squared etc.) |
| Performance metrics (ME, MER, RMSE, N-RMSE) | Comparing patient sequence error w.r.t standard template | Over-fitting, difficult to generalise | Statistical analysis (ANOVA, Chi-Squared etc.) | |
| Statistical analysis | Tests to see and compare different results | Comparison of kinematic parameters | Cannot categorise or grade patients | Time series comparison algorithms such as HMM |
| POMDP | MDP where underlying process is not directly observable | virtual game assessment | Intractable, assumes convex value functions | Reinforcement learning |
| KNN | non parametric classification algorithm | classification of kinematic parameters | Need to determine K value, high computation cost | SVM |
| SVM | Classifies by finding discriminating hyperplanes | Classification of kinematic parameters | choosing appropriate kernel, memory intensive | Ensamble, RF methods, Neural networks |
| Neural fuzzy system | Learns fuzzy parameters with neural networks | Builds model-based on clinician experience instead of rules | Difficult to interpret results and generalise | DL models |

| | | | | |
|---|---|---|---|---|
| ANN | Stochastic learning networks that learns from examples | Classification of kinematic parameters | Prone to over-fitting, poor performance with images or sequential data | CNN, LSTM, TCN |
| GRBM | RBM that uses binary latent variables to model hidden states | Pose recognition | Difficult to train with contrastive divergences, rRquires sampling like GS, MCMC | MRF, CRF, AE |
| MDNN | Combination of ANN and Mixture density models | Classification through log-likelihood based performance metrics | Number of mixture modes need to be fixed manually | Various adaptations of MDNN (Makansi et al., 2019; Wang et al., 2019; Ye; Kim, 2018) |
| HMM, HSMM | Represents probability distributions over sequences | Discriminate between patient and model skeleton sequences | Limited by Markov property, cannot capture higher order dependencies | CRF, Bayesian Networks |
| DTW, MDDTW, SS-DTW | Measures similarity between two temporal sequences | Calculate similarity between patient and model skeleton sequences | Quadratic complexity, Works with only smaller templates | HMM |
| CNN | Networks that learns grid like topology like images | Classifying poses from RGB images, videos | Resource consuming, requires big datasets | GNN (Kipf; Welling, 2016), Capsule Networks (Sabour et al., 2017) |
| MSNB | Bayesian classifier variable independence assumption | Classification of kinematic parameters | Requires explicit modelling of inter-dependence between variables | Fuzzy logic, different types of SNB (Zheng; Webb, 2005) |
| LSTM | RNN architecture, learns temporal sequences | Learning skeleton sequence features | Slow, cannot memorise long temporal sequences | TCN (Lea et al., 2017), ODE networks (Chen et al., 2018a) |

Table 2.9: A summary of feature comparison methods used, their drawbacks and alternatives that can be used. CRF: Conditional Random Fields, GNN: Graph Neural Networks, MRF: Markov Random Fields, ODE: Ordinary Differential Equation, TCN: Temporal Convolutional Networks

In Table 2.9 various feature comparison methods used by authors are highlighted along with drawbacks and possible alternatives. Most basic methods used by researchers are simple numerical and graphical comparison (Fern'ndez-Baena et al., 2012; Exell et al., 2013) of skeleton trajectories, joint angles or other kinematic parameters. The results are hard to generalise beyond the examples presented and may lack statistical significance. A better alternative is to use statistical tests such as ANOVA (Kurillo et al., 2013), Chi-Squared tests, Co-relation methods and so on. Graph trajectories can be compared with methods such as Kullback Leibler Divergence (KLD) which could provide statistically significant results. Authors have also used distance measures such as Euclidean Distance (Antunes et al., 2016), Mahalanobis distance (Liao et al., 2019) for comparing features. Model-based approaches perform better than the statistical approaches mentioned so far which are model-less approaches (Bishop, 2006). Thus, authors (Palma et al., 2016b; Capecci et al., 2016) have used temporal sequence comparison algorithms like HMM, DTW or their variants such as HSMM MD-DTW and so on with distance-based objective function for sequence comparison. However, other authors(Taati et al., 2012; Leightley et al., 2013) have used classification algorithms such as K-means, SVM to classify modelled sequences through HMM, DTW etc and Taati et al. (2012) show that combination of modelling and classification algorithm works better than using either alone. Sequence comparison can be also carried out with techniques such as CRF, or through generative models like Boltzmann Machines or Bayesian Networks. These statistical algorithms have largely been replaced

by DL-based algorithms such as CNN, LSTM and TCN (Oyedotun; Khashman, 2017; Alom et al., 2018; Liao et al., 2019). When comparing sequential data with DL, LSTM is the most popular type of architecture that has been used. But recently TCN (Lea et al., 2017) and ODE networks (Chen et al., 2018a) have shown very competitive results and these two architectures are being actively pursued by researchers. CNNs have been almost exclusively used for processing image and video data but authors in this domain are yet to take advantage of the widely used state-of-the-art architectures (Aloysius; Geetha, 2017). Researchers in the broader CV community are now exploring the use of more advanced techniques including, but not limited to Capsule Networks (Sabour et al., 2017), effective scaling networks such as MobileNets (Howard et al., 2017) Efficient-Nets (Tan; Le, 2019), GNNs (Kipf; Welling, 2016). Authors in this domain can adapt some of these state-of-the-art techniques for performance improvement.

## 2.11 Discussion and Conclusion

The review was the first step undertaken as a part of the current study. It helped to understand and provide the motivation to explore some of the gaps in the current literature. The discussion highlights some of these gaps and relates to the stated objectives of the current study. The near absence of image, video-based DL algorithm is quite contrasting to other areas such as pose estimation and action recognition where such algorithms have been widely used. In this domain, most articles exclusively use skeletal information as raw data. This means images, low-level image/video features and high-level contextual cues (e.g., body-objects interaction) are not part of the intelligent processing. This can lead to loss of valuable contextual information. This prompts the current work to further the research on video-based activity recognition (Chapter 6), combined video and pose-based activity recognition model (Chapter 6). The review also shows the area suffers from the lack of publicly available datasets which is vital for the involvement of modern highly successful and data-driven DL-based models. As explained in Chapter 1 (Sec. 1.1.2) the lack of datasets could be a major reason behind the absence of DL-based research in this domain. Further, the review shows that existing research has approached this domain in various ways but have not explored CV-based functional assessment of physically impaired persons using ADL. Functional assessment through ADL is widely carried out for assessing a patient's condition and various methods have been proposed to measure the same (Green; Young, 2001). This provides the motivation for preparing a dataset that captures different physical impairment-specific versions for 10 different ADL (Chapter 5). Motivated by the comparative absence of DL-based models in this domain, the study presents a DL-based multi-label ADL recognition model that can discriminate between various impairment specific versions of the same ADL. In this way, the research aims to contribute towards automating the assessment of physically impaired persons through ADL.

The review presents a well summarised and analysed collection of major CV-based research in rehabilitation and assessment of persons with physical impairments. It proposes its own taxonomy based on end-user application. To the best of my knowledge, this is the only review article to date, that covers the latest advances in this application area and presents them from a CV application point of view. It particularly focuses on the type of data, feature representation and comparison algorithms

employed by authors to assess physically impaired persons. Reviewing and analysing the comparison techniques is especially important due to the wide ranging and hugely varying manifestations of abnormal or impaired human movements. Owing to the varying nature of human body motion and its impairments, research in this area has involved vastly varying techniques. These range from simple graphical comparison of joint angle trajectories to application of complex algorithms such as GANs. The review has been accepted for publication in *Springer Multimedia Systems* after three reviews. Owing to the the lack DL-based methods in the articles reviewed, the next Chapter presents a review of DL-methods relevant to the current study.

# Chapter 3

# Literature Review: Deep Learning

## 3.1 Introduction

This Chapter presents a literature review relevant to the proposed DL approaches for human activity recognition and pose-estimation. The previous Chapter presented a review of CV-based approaches for automated assessment and rehabilitation of physically impaired persons. The review highlighted that research in this domain is yet to fully explore AI or DL-based approaches. The current study aims to contribute towards this domain by exploring DL-based approaches for recognising physical impairment-specific ADL (Chapter 1, Figure 1.2). Thus, this Chapter presents DL-based literature relevant to the models proposed in this study. There are many aspects to modern DL-based models and as shown in Figure 1.2 (Chapter 1), the study focuses on four such aspects. These include efficient spatial and temporal processing of human activity video and body-pose sequences. Accordingly, this Chapter presents widely used standard DL architectures and techniques for spatial temporal processing such as CNN, LSTM and TCN. The study further exploits two widely used techniques called 'Attention' mechanism and 'Pooling'. The proposed human activity recognition and pose-estimation models incorporate 'Attention Mechanism' and 'Pooling' in a novel way to enhance the models' recognition/estimation accuracy. Then, the literature relevant to human pose estimation and human activity recognition is discussed which forms the basis of the models presented in the current study. The review on pose-estimation model focuses on the motivation to design a lightweight pose estimation model. The human activity recognition review is divided into video-based models and pose-based models. It highlights the motivation for designing a novel joint position encoding algorithm method prepared for the pose-based models presented in this study.

## 3.2 Artificial Neural Networks

### 3.2.1 Spatial Processing Networks

The introduction of LeNet (LeCun et al., 1998) for document recognition revolutionised the field of AI and put ANN at the forefront of AI-based research. However, their architecture which consisted of two convolutional layers and an FC layer was inadequate for large-scale visual recognition tasks. Thus, the well-known ImageNet challenge (Krizhevsky et al., 2012) which forms the benchmark for large-scale visual recognition remained dominated by classic CV-based methods until 2012. In 2012, AlexNet (Krizhevsky et al., 2012) introduced CNNs for image-classification task in their

seminal research and won the 2012 ImageNet challenge for image classification. Since then, CNNs, which are a type of ANN, have been almost exclusively used for image processing tasks such as object detection, object classification, human activity recognition, human pose estimation and so on. AlexNet, which was a much-improved version of LeNet (LeCun et al., 1998) consisted of five convolutional layers followed by two FC layers. The additional representations produced by the extra layers greatly increased the learning capacity of AlexNet (Krizhevsky et al., 2012) as compared to LeNet (LeCun et al., 1998). However, the additional layers massively increased the number of parameters in AlexNet which meant the network required more processing power and was prone to over-fitting (Goodfellow et al., 2016). After AlexNet, ImageNet dataset saw increased accuracy with a succession of CNN-based architecture such as VGG (Simonyan; Zisserman, 2015), Inception (Szegedy et al., 2015), ResNet (He et al., 2016) and Inception-ResNet-V2 (Szegedy et al., 2017). Szegedy et al. (2015) realised just adding layers did not increase performance but increased over-fitting. Instead, the authors proposed 'Inception' modules that effectively captured spatial representations from images with a reduced number of connections between individual neurons. The 'Inception' modules split a convolution operation into a two-layered operation which reduced the number of connections. He et al. (2016) demonstrated that adding residual connections remarkably increased the performance of CNNs. Residual connections worked by propagating high-level information captured by earlier layers to the top-layers, which otherwise were lost due to weak gradient flow (He et al., 2016). Szegedy et al. (2017) combined inception modules with residual connections for the Inception-ResNet-V2 model. Inception-ResNet-V2, which is a high-performance image classification model, has been adapted for video-based activity recognition models presented in this study.

Training aforementioned deep networks are expensive in terms of resources (e.g., GPUs) and computational complexities. Transfer learning from pre-trained networks was introduced as a way to overcome the drawback to some extent. Yosinski et al. (2014) explained that generalised or high-level features from the first few layers of a pre-trained network were transferable. The current consensus is that layers towards the end learn composite features specific to a given task. Whereas the layers towards the beginning tend to learn more general features such as edges, corners and so on. The trend in image classification shifted to use ImageNet (Krizhevsky et al., 2012) pre-trained networks and apply transfer learning on the target datasets which made the training faster and computationally less demanding. Recently, Zoph et al. (2018) proposed an advanced transfer learning method through Network Architecture Search (NAS) model. NasNet searched for the best performing models by searching for well-trained layers in a small dataset and transferred the layers in a new architecture for a larger dataset. In Deep-cut (Pishchulin et al., 2016), ImageNet pre-trained network was used to fine-tune the pose estimation model. Well-known models such as Open-pose (Cao et al., 2018) used pre-trained VGG-16 (Simonyan; Zisserman, 2014) for feature extraction which had 133 to 144 million parameters and this added to the computational cost of inference. Therefore, inference using such networks are not suitable for mobile-based pose estimation, where the use of high-performance GPUs is not feasible. Assessment of physically impaired persons at home or in clinic is an example where a lightweight model may be more suitable. Exploring DL models for mobile-based systems is an active area of research where the trade-off is getting a good balance between speed and accuracy. One such model, SqeezeNet matched AlexNet's (Krizhevsky et al., 2012) performance with 50 times

fewer parameters (Iandola et al., 2016). It reduced the number of parameters by replacing 3x3 filters with 1x1 filters. It also maintained large activation maps by downsampling late in the network which increased the size of maps and hence accuracy. However, despite the large activation maps, the overall size of the model was small due to the requirement of fewer filters. DenseNet (Huang et al., 2017) showed that if every layer was connected to every subsequent layer in a residual manner, the network achieved similar accuracy with less parameters. The authors demonstrated that, to achieve similar accuracy to that of ResNet-152 (He et al., 2016), DenseNet required less than half of the parameters. ResNet (He et al., 2016) was the first to connect different layers with residual connections however the residual connections were relatively sparse as compared to DensNet (Huang et al., 2017). Huang et al. (2017) showed that by connecting every layer to all of its subsequent layers in a block, DenseNet achieved a much better gradient flow as compared to ResNet (He et al., 2016). Xception (Chollet, 2017) used depth-wise separable convolutions for constructing a lighter model. MobileNets (Howard et al., 2017) combined depth-wise separable convolutions with point-wise separable convolutions for designing an architecture with only 4.3 million parameters. The factoring of a convolutional operation into depth-wise and point-wise separable convolutions greatly reduced the number of parameters which made the network faster. In this study, the research on pose estimation focuses on adapting MobileNets for pose estimation and it takes advantage of transfer learning for quick training with less data. Efficient-Net architecture relied on multi-dimensional (depth, width, resolution) scaling for a group of networks that performs across the speed vs accuracy spectrum (Tan; Le, 2019). The lightest model EfficientNet-B0 has around 3 million parameter and is the best performing model on the ImageNet dataset given its size. On the other hand EfficientNets-B7 with 70 million parameters achieved the highest top-1 accuracy with ImageNet.

### 3.2.2 Temporal Processing Networks

Human activity data is nothing, but sequence of frames composed of RGB video, depth or human body-pose information. Thus, human activity recognition requires modelling of temporal information and dependencies between frames. RNNs are a type of ANN that are widely used to model temporal data (Goodfellow et al., 2016). RNNs work by forming directed graph connections between units along the temporal dimension. This allows the network to capture the temporal dependencies of a time series or sequence. Unlike normal feed-forward networks, RNNs have internal memory states that are used to process temporal inputs. For long temporal sequences, RNNs suffer from 'Short-Term' memory problems. If the temporal sequence is long, RNNs fail to carry information from one end of the temporal sequence to the other (Hochreiter; Schmidhuber, 1997). RNNs also suffer from vanishing gradient (Hochreiter, 1998) problem during back-propagation through time, which means that the gradients become too small to contribute to learning. Hochreiter; Schmidhuber (1997) introduced LSTM networks to overcome these drawbacks. LSTM cells consisted of forget, input and output gates which act as a memory for preserving long temporal information. Thus LSTMs mitigated the 'Short-Term' memory problems of simple RNNs. Learning introduced by LSTM memory also helped to mitigate the vanishing gradient problem. Gated Recurrent Units (GRUs) were introduced as newer and simplified version of LSTM which did not use the output gate (Cho et al., 2014). GRUs together with LSTM are almost exclusively used for temporal processing

involving RNNs. Human activity recognition involving skeleton pose, video or both are essentially sequential data and authors have extensively exploited LSTMs for temporal processing (Liu et al., 2016b; Li et al., 2017b; Aliakbarian et al., 2017; Song et al., 2016). Inspired by this, the LSTM has been used for temporal processing of both RGB and pose data in this study (Chapter 6 and 7). The study used LSTM in an 'Attention'-focused (Sec. 3.2.4) manner through the use of a 'Self-Attention' mechanism that enhanced the model's discriminatory capability. In the RGB video-based model (Chapter 6) the hidden LSTM states have been semantically clustered in a novel manner to enhance the model performance further.

Recently, Lea et al. (2017) introduced TCN based on 1D convolutions to process time-series data for action detection and segmentation. Their encoder-decoder TCN (ED-TCN) was based on the well-known Wavenet TCN model (Oord et al., 2016) for generating raw audio waveforms. Similar to an Auto Encoder (AE), encoder part of ED-TCN down-samples and the decoder part up-samples the input. It used a deep stack of dilated convolutions to capture long temporal patterns. Causality was maintained, as weights taking input from a particular frame, were only connected to other weights that received input from past frames. In recurrent networks (LSTM or GRU) the hidden state at each temporal point $t$ was only a function of the input at $t$ and the previous hidden state that represented the input at $t-1$ (Goodfellow et al., 2016). The authors (Lea et al., 2017) argued that this sometimes limited the capacity of LSTMs to model long-term temporal dependencies. TCNs overcame the limitations of GRUs and LSTMs by efficiently capturing long-term temporal dependencies through dilated one dimensional (1D) convolutions (Lea et al., 2017). In dilated convolution, a filter was applied over a length which was larger than the filter length (Lea et al., 2017). This resulted in an increased size of the receptive field without increase in computation costs. Thus, the computations were faster along with more effective representations of long-range temporal patterns. Kim; Reiter (2017) combined the concept of TCN with residual connections (He et al., 2016) to propose a pose-based model for activity recognition. In the current study, TCN has been used with a novel human pose encoding method for pose-based human activity recognition (Chapter 7). The efficacy of the novel pose encoding method is further demonstrated in Chapter 8 by integrating the encodings to TCN-based network introduced by Kim; Reiter (2017).

### 3.2.3 Activity Recognition - Learn-able Pooling

Typically, the final layer of a neural network used to be a FC layer (LeCun et al., 1998; Krizhevsky et al., 2012). Directing the output of convolutional or RNN layers to FC layers caused a massive increase in the number of parameters. This made the network computationally more expensive and prone to over-fitting (Goodfellow et al., 2016). To prevent over-fitting and reduce the number of parameters in the FC layer, often a pooling layer is used to downsize the convolutional or RNN maps (Howard et al., 2017; Kim; Reiter, 2017). There are various pooling mechanisms in literature like Average or Max Pooling (Habibian et al., 2016; Hussein et al., 2017), Rank-Pooling (Fernando et al., 2016), Context-Gating (Miech et al., 2017) and High-Dimensional Feature encoding (Xu et al., 2015b). In Global Max Pooling (GMP), the maximum value in each of the CNN maps is passed onto the FC layer whereas in GAP, the average value is considered. However, pooling using statistical

methods or high dimensional encoding does not select the more important or discriminatory features contained in the maps produced by CNN or RNN. Thus, authors have explored learn-able pooling methods to pool the most relevant features based on learned representations. Girdhar; Ramanan (2017) introduced 'Second-Order Attentional' pooling in which the final output map from CNN was multiplied with a weighted version of itself. The weights learn representations that guide the network to pool more important discriminative features instead of pooling using statistical methods. However, second-order pooling led to increased values in maps which required Girdhar; Ramanan (2017) to use rank-1 approximation techniques to avoid computing second-order features. A low-rank approximation is not always accurate and normally various techniques are required to improve accuracy (Kishore Kumar; Schneider, 2017). Researchers have also approached learn-able pooling by integrating well-known image feature descriptors with DL-based models. (Arandjelovic et al., 2016; Miech et al., 2017). Image feature descriptors such as Vector of Laterally Aggregated Descriptors (VLAD) (Jégou et al., 2010) and FV (Perronnin; Dance, 2007) rely on aggregation of unsupervised clustering information. VLAD used K-means, while FVs used GMM for clustering. Arandjelovic et al. (2016) introduced NetVLAD, where VLAD clusters were learnt in a supervised manner and used as input for learn-able pooling mechanism towards the end of the network. In a similar manner, Girdhar et al. (2017) introduced Action VLAD, where VLAD features were used as input for learn-able pooling for activity recognition. NetVLAD and Action VLAD integrate K-means with CNN based models to learn semantic clusters in a semi-supervised manner. K-means is a hard assignment clustering mechanism which means the data points are fully assigned to a single group (Bishop, 2006). On the other hand GMM assigns the data points in a soft-assignment form meaning a data point can be assigned to multiple clusters where the weightage of all cluster assignments sum up to one. Soft-assignment is considered to be more flexible than hard assignments and Smith; Kornelson (2013) showed that FV-based on GMM performed better than VLAD based on K-means. FV was the first and second-order aggregation of cluster weights, cluster means and co-variances (Perronnin; Dance, 2007) of a GMM. Miech et al. (2017) introduced learn-able FV (NetFV) to semantically cluster and pool audio and video features, where FVs were learned through the DL model. Unlike in original FVs, the cluster weights in NetFV were not calculated from GMM but were calculated using a differentiable soft-assignment (Miech et al., 2017). The current study explores intelligent pooling based on FV inspired by the success of intelligent pooling methods based on 'learn-able' mechanism in contrast to statistical methods. In Chapter 6, NetFV (Miech et al., 2017) is adapted in a novel manner to semantically cluster the temporal structures and relationships in hidden Bi-LSTM states. The semantically clustered states are pooled in an activity-aware manner which obviates the need for further processing through FC layer thereby minimising over-fitting. The effectiveness of the proposed method is further demonstrated in Chapter 8 where the FV-based method is used to intelligently pool maps produced by TCN-ResNet (Kim; Reiter, 2017).

### 3.2.4 Attention mechanism

Inspired by human visual search mechanism, Bahdanau et al. (2014) introduced 'Attention' mechanism to ANNs to selectively focus on more relevant and discriminatory features. The mechanism calculated the similarity between the input vectors 'queries' and 'keys'. This similarity measure is

called 'context vector' and it represents weights that are to be applied to input vector 'values' to represent a weighted version of 'values' to the output (Bahdanau et al., 2014). Thus, 'Attention' mechanism maps the 'values' weighted with 'context vector' to the output. Often, 'keys' and 'values' are the same vector. There are several mechanisms to calculate the 'context vector'. This includes, application of learn-able weights to 'keys' and 'values' and calculating either of 'Additive-Attention' (Bahdanau et al., 2014), 'Dot-Product Attention' (Luong et al., 2015), 'Scaled Dot-Product Attention' (Vaswani et al., 2017) and so on. Soft-max activation function is applied to the result so that the final 'Attention' map adds up to one. Zhang et al. (2018) proposed a 'Self-Attention' mechanism that relates various temporal positions of the same sequence to calculate a weighted representation of itself. Therefore, in 'Self-Attention', 'queries', 'keys' and 'values' are the same vector. 'Self-Attention' has been shown to be very effective in many tasks including, but not limited to abstractive summarisation, machine reading and image description generation (Cheng et al., 2016; Zhang et al., 2018). However, Vaswani et al. (2017) argued that capturing long-term temporal dependencies and structures becomes difficult with a single representation for long sequences. Thus, Vaswani et al. (2017) introduced 'Multi-Head Attention' mechanism that linearly juxtaposed the output of 'Scaled Dot-Product Attention' into a number of groups (heads). The authors demonstrated that this allowed the model to represent different learned sub-spaces at different positions. By enhancing the number of representations from a single representation (Self-Attention) to multiple heads, Vaswani et al. (2017) achieved better performance. However, the increase of learned sub-spaces increased the number of parameters making 'Multi-Head Attention' (Vaswani et al., 2017) susceptible to over-fitting as compared to 'Self-Attention' (Zhang et al., 2018).

'Attention' mechanisms have been widely adopted for image and video understanding tasks with spatial, temporal 'Attention' and spatio-temporal variations (Cho et al., 2015; Xu et al., 2015a; Sharma et al., 2016; Song et al., 2017; Jaderberg et al., 2015). Xu et al. (2015a) proposed 'Hard' and 'Soft Attention'-based visual 'Attention' models for image captioning. The 'Hard-Attention' is a stochastic technique based on Monte Carlo methods, whereas the 'Soft-Attention' is a deterministic technique where the context vector is determined by optimising the marginal log-likelihood to calculate the expected value (Xu et al., 2015a). Song et al. (2017) proposed an end-to-end spatial and temporal 'Attention' network from human body pose features. The separate spatial and temporal 'Attention' sub-networks based on LSTMs aided the main network to pay different levels of attention to each frame which enhanced the discriminatory capability of the main network. Sharma et al. (2016) introduced a Soft Attention-based model with LSTM that learnt to pay attention to parts of a frame after taking few glimpses. Matsuo et al. (2014) proposed 'Regional Attention', which focused the network on objects important in a region of the image for ego-centric activity recognition. Sudhakaran et al. (2019) proposed a Long Short-Term Attention network that focuses on features from relevant spatial parts while 'Attention' is tracked smoothly across the video sequence for ego-centric activity recognition. Song et al. (2017), Sudhakaran et al. (2019), and Sharma et al. (2016) have focused on improving spatial processing of CNN maps through the integration of 'Attention' mechanism to LSTM. This helped LSTM to focus on more discriminatory features and processed the long-term temporal dependencies more effectively. However, it is not clear why the authors (Song et al., 2017; Sudhakaran et al., 2019; Sharma et al., 2016) used 'Attention' mechanism with LSTMs which are good at temporal processing, to improve the spatial representation of maps produced by CNN. In

contrast, the current study explores the application of 'Attention' mechanism (Multi-Head and Self) to focus the visual features extracted from deep CNN for improving the discriminatory capabilities of the model. These 'Attention'-focused CNN maps are then used with LSTM to improve the temporal processing capabilities of the video-based models presented in this study.

## 3.3 Human Pose Estimation

Research focus on human pose estimation has shifted from classical approaches (Li et al., 2008; Ramakrishna et al., 2014; Yang; Ramanan, 2011) to deep neural networks since the introduction of AlexNet for pose estimation (Toshev; Szegedy, 2014). Human pose estimation is a regression problem and stacked hourglass network (Newell et al., 2016) has become the basis of many pose estimation models (Ning et al., 2017b; Yang et al., 2017; Ke et al., 2018). The stacked hourglass network goes from high resolution to low and back to high and hence the name hourglass. It also has skip connections that connect the downsampling layers to the upsampling layers. The high output resolution made it suitable for heat-map regression. Models based on the stacked hourglass network have included hand-crafted features to guide their network better (Ning et al., 2017b; Yang et al., 2017). Cao et al. (2016b) designed their own network using 'Part Affinity Fields' for multi-person pose estimation. Supervision is carried out using heat-maps where each heat-map represent a single joint of all the persons. Joints for each person are then associated through the 'Part Affinity Fields'. Pose estimation problem has also been addressed as a body part classification and joint localisation problem. The Deepcut model (Pishchulin et al., 2016) used a partitioning and labelling formulation generated with CNN-based part detectors. Use of R-CNN-based classification to achieve joint localisation has also been explored in Gkioxari et al. (2014). More recently, researchers have explored adaptation of standard CNN-based classification architectures for pose estimation. Xiao; Wan (2017), used ResNet-50 for progressively regressing the joint coordinates. Papandreou et al. (2017b) used ResNet architecture with faster R-CNN and regressed heat-maps in a novel way of non-maxima suppression. Combination of concepts from pose estimation and classification has also been used together. Ning et al. (2017a) combined modules from Inception-ResNet (Szegedy et al., 2017) within stacked hourglass network along with hand-crafted features such as HoG and Hough features for multi-person pose estimation. AlphaPose is based on Regional Multi-Person Pose Estimation (Fang et al., 2017), which minimised the inaccuracies in the bounding box and redundant human detection using Symmetric Spatial Transformer Network, Parametric Pose Non-Maximum-Suppression and Pose-Guided Proposals Generator. For better performance AlphaPose model has now been trained on Crowd-Pose dataset (Li et al., 2018a) in addition to the original MSCOCO dataset. The DL models mentioned have achieved very high accuracy for 2D human pose-estimation. However, these models use very high-performance GPUs for inference which make them infeasible for lightweight pose estimation scenario. For example, Cao et al. (2016b) uses a pre-trained VGG-16 (Simonyan; Zisserman, 2014) which has around 144 million parameters and require high performance GPUs for inference. In contrast, the current study uses the well-known lightweight classification model MobileNets (Howard et al., 2017), which has only 4 million parameters for preparing a lightweight pose estimation model. In the current study a pre-trained MobileNets is adapted to resemble the highly successful stacked hourglass (Newell et al., 2016) architecture.

2D human pose estimation has been extensively explored by the computer visionCV community. However, in areas like pose-based human activity recognition, the recent best performing models have mostly relied on 3D pose estimation. Typically, a depth sensor like Kinect (Han et al., 2013) is deployed to capture 3D pose information along with RGB and depth images. For the proposed dataset Kinect has been used to capture the RGB, depth and pose information. However, Kinect is based on old technology and suffers from inaccuracies as highlighted by Galna et al. (2014). The success of 2D pose estimation has led authors to explore DL for 3D human pose estimation from monocular images (Chen; Ramanan, 2017; Martinez et al., 2017; Nie et al., 2017). These are two-step approaches where the first stage carries out 2D pose-estimations and the second stage regresses depth estimates of the 2D coordinates predicted in the first stage. Yang et al. (2018) used GAN to predict 3D pose from monocular images in the wild. In the absence of ground truth for 3D pose information from monocular images, authors used semi-supervised methods for 3D pose estimation. Pavlakos et al. (2018) used weak supervision through ordinal depth relations (closer-farther) of the human body to regress their hourglass-like network. Pavllo et al. (2019) used semi-supervised learning with TCN for 3D pose estimation. Kocabas et al. (2020) used a temporal GAN model to estimate human body shape. However, 3D pose-estimation from monocular images is still largely experimental and as seen from previous Chapter (Chapter 2), authors have mostly used Kinect which relies on depth data for 3D human pose estimation.

## 3.4 Human activity recognition

Human activity recognition is a major objective of this study and in this section a short overview of relevant literature is presented. First, classical vision-based approaches are reviewed and then modern AI-based techniques are presented.

### 3.4.1 Classical Approaches

According to the human motion tracking survey by Moeslund; Granum (2001), action recognition can be classified into two paradigms: recognition by reconstruction and direct recognition. Many of the early works in this area relied on direct recognition with robust low-level features. Polana; Nelson (1994a) used spatio-temporal templates of motion features for matching against low-level features. Davis; Bobick (1997) used temporal templates for comparison using Mahalanobis distance. The authors (Polana; Nelson, 1994a; Davis; Bobick, 1997) discriminated activities by employing model-less statistical techniques (Mahalanobis distance, Template matching) which lacked generalisation, were not-scalable and have been proven inferior to model-based techniques (Bishop, 2006). Thus, model-based techniques as mentioned next emerged as better alternatives. The seminal work of Yamato et al. (1992) relied on binary frames type human area extraction from the video as input data for their HMM based model. Aggarwal (2004) built a Deep Bayesian Network model for a 3-tier recognition system. This was one of the very first works to recognise action from human body parts rather than primitive information such as silhouette. Heisele; Woehler (1998) used combined colour and position based feature space to depict pedestrians. The research then built a two-stage classifier. The first

stage was a fast polynomial classifier and the second stage was a time-delay neural network classifier to recognise the walking action of pedestrians. Colour-based features are prone to background noise, illumination and thus optical flow-based features (Lucas; Kanade, 1981) as used by Ullah et al. (2018) maybe more suitable for intelligent processing of video data. The research by Bregler (1997) was another important work in this area which relied on multiple staged data extraction and classification for gait recognition. The novelty of the work lies in the incorporation of past histories of groupings in earlier frames and encoding of spatial proximity. It used gradients and textures for coherent blob detection and a sequential combination of HMM, Kalman filter, a mixture of Gaussian and clustering, to achieve the goal. However, it is not clear why Bregler (1997) used unsupervised clustering through expectation maximisation. Normally, for labelled data (human activity videos) supervised classification techniques such as SVM are considered more appropriate (Bishop, 2006) and are widely used (Lublinerman et al., 2006). Chomat; Crowley (1998) used spatio-temporal features refined by PCA along with Bayes classifier for discrimination between actions. Masoud; Papanikolopoulos (2003) also used PCA to bring down the dimensionality for efficient computing. It compared manifold of actions in Eigen-space against a pre-defined manifold for action recognition. PCA used in the last two methods is a linear dimensionality reduction technique which may not be the best option for the highly non-linear human motion. As seen from the previous Chapter (Chapter 2, Sec. 2.7) non-linear techniques such as non-linear PCA, diffusion maps and others may be more suitable in such cases.

The methods discussed so far were largely based on generative models where a pre-defined model was generated for comparison. As with human pose estimation, human action recognition began to take advantage of supervised discriminative machine learning algorithms (Batra et al., 2008; Lublinerman et al., 2006; Xia et al., 2012a). However, these algorithms relied on better human motion tracking methods than simple low-level tracking. Wang et al. (2006b) used canny edges to extract human silhouettes and fed them into a spectral clustering (Ng et al., 2002) algorithm for unsupervised action recognition. Similar to colour-based features silhouette-based features suffered from background noise, varied illumination and thus were largely replaced by more advanced mid-level features (Batra et al., 2008; Gorelick et al., 2007). Batra et al. (2008) built a dictionary of code-words from mid-level features called space-time shape-lets and used it with KNN classification. The dictionary of code-words was an example of a data-oriented approach whereas the earlier approaches of human tracking were semantic approach. Gorelick et al. (2007) utilised Poisson equation solution properties to get space-time features that were better suited for classification. It used nearest-neighbour with Euclidean distance on normalised global features for classification. Hierarchical activity recognition also came into play around this time. Human activity can be viewed as composition of multiple levels of low-level activities arranged in a hierarchy which leads to the overall activity (Ryoo; Aggarwal, 2009). Thus, looking for low-level activities and by repeatedly applying the recognition algorithm one can determine the overall activity. Ryoo; Aggarwal (2009) designed a spatial-temporal kernel to hierarchically recognise up to six different activities in a multiple subject scene. Lublinerman et al. (2006) used Fourier descriptors to represent the human silhouettes and fed them to SVM classifier for action recognition. This is another example of a data-driven model followed by a discriminative classifier. Kovashka; Grauman (2010) used bag of words but learned candidate neighbourhoods to build the most informative configurations. These descriptors were recursively mapped to higher-level

vocabularies which is another example of hierarchical recognition. Sempena et al. (2011) used DTW to compare human pose sequence against a pre-defined sample for action recognition. Although human pose estimation and other classification tasks relied more on discriminative classification, due to high dependency on temporal data, generative HMM-based models retained their significance. Xia et al. (2012a) used Kinect to extract 3D joint locations and clustered them into visual words, but instead of a discriminative classifier, it used generative HMM. It showed that precise joint localisation increased the robustness of models. Wu et al. (2014) used Kinect to localise pose and used SVM and HMM together for action recognition. Jalal et al. (2017) used extensively processed depth maps-based silhouettes with HMM to perform action recognition for healthcare monitoring systems. The authors (Jalal et al., 2017) demonstrated that unlike silhouette based on RGB image, depth-based silhouettes were much more robust to noise, missing joints and captured local dependencies in a more effective manner.

In later works, the focus of research shifted from low-level features and generative models to more CV-based feature descriptors with discriminative classifiers (Behera et al., 2014; Peng et al., 2014; Shahroudy et al., 2015; Peng et al., 2014). Behera et al. (2014) used SIFT based spatio-temporal features to build a RF model with a discriminative Markov decision tree algorithm for ego-centric action recognition. Peng et al. (2014) relied on multiple feature space representations using HoG, Histogram of Optical Flow (HOF) and Motion Boundary Histogram (MBH) with SVM classifier for action recognition. Lan et al. (2015) enhanced Peng et al. (2014) with a multi-skip feature tracking. Shahroudy et al. (2015) combined dense trajectories consisting of HoG, HOF and MBH alongside depth information. CV-based feature descriptors such as SIFT, MBH are more robust to variation in scale, background noise illumination and so on (Forsyth; Ponce, 2012). SIFT used by Behera et al. (2014) is a robust well-known scale-invariant, rotation-invariant local feature descriptor. It is more robust to background noise and illumination than colour-based features or silhouette extraction from morphological operations (Forsyth; Ponce, 2012; Lowe, 2004). On the other hand, MBH, HOF are motion-based descriptors that depend on optical flow (Dalal et al., 2006), for calculating robust temporal descriptors. Chen et al. (2016) explored maximum likelihood estimation for classifying their action graph consisting of 3D pose-based motion features. Luvizon et al. (2016) extracted displacement vectors from skeletal sequences, then feature aggregation by K-means clustering, PCA and VLAD. This was followed by multiple stages of KNN based classifiers for combining similar features before classifying an action. The authors show that KNN worked better than SVM or ANN for their model. However, with only two FC layers the model used by the authors was very basic as compared to other DL-based activity recognition models (Wang et al., 2012). Uddin et al. (2017) combined HOF and local ternary pattern to form adaptive rich feature description from simple background subtraction. The model effectively combined flow feature descriptor HOF with appearance descriptor based on extension of ternary pattern used for static texture analysis.

### 3.4.2 Deep Learning Approaches

In the last decade, researchers began to exploit DL approaches based on spatial processing networks CNN and temporal networks such as LSTM, TCN for human activity recognition. CNN-based architectures have been highly successful for tasks that require spatial processing such as image recognition

and so on. On the other hand, temporal networks are used for sequential processing tasks such as natural language processing, speech recognition and so on. Human activity recognition requires both spatial and temporal processing and as described in the following discussion, authors have exploited both for activity recognition.

#### RGB video-based models

Traditional approaches, involving both classical and DL techniques have mainly focused on monocular RGB video data (Herath et al., 2017). Authors first started using classical vision-based features for spatial processing followed by LSTM for temporal processing. Baccouche et al. (2010) were one of the first to use LSTM for video action classification where the video features were represented by SIFT in Bag of Visual Words (BoVW). Grushin et al. (2013) implemented LSTM with HOF features. Then, with the success of CNNs the focus shifted to the use of CNNs for spatial processing instead of simple features like SIFT, BoVW, HOF and so on. CNNs offer richer feature description via learning as compared to classical hand-crafted CV-based features and thus perform better (Goodfellow et al., 2016). Broadly, there are two ways to represent video data through CNNs. In the first method, stacked sequences of frames are encoded through 2D CNN in a time distributed manner where weights of CNN are shared across frames (Ma et al., 2016; Deng et al., 2016). This is typically followed by a temporal processing network consisting of LSTMs (Donahue et al., 2016). In the second method the RGB video data is directly fed into a 3D CNN (Wu et al., 2016a; Molchanov et al., 2016; Ji et al., 2012). Tran et al. (2015) introduced the concepts of 3D CNNs where all the frames from a sequence of frames were processed together as a single input. Molchanov et al. (2016) introduced recurrent 3D CNN for online detection of hand gestures. Carreira; Zisserman (2017) proposed the two-stream I3D architecture using inflated 3D CNNn. In inflated 3D CNN the filters and pooling kernels of the network are expanded into 3D. The enables the network to learn seamless spatio-temporal feature extractors from video. 3D CNNs are beneficial as they take advantage of image processing capabilities of CNNs and add temporal processing capability. However, 3D CNNs need a lot of computational resources as these networks simultaneously process spatial and temporal dimensions. On the other hand, 2D CNNs only process one frame at a time. Thus, to process long-term temporal sequences with 3D CNNs, frames have to be pooled. This results in loss of valuable temporal information (Singh et al., 2016).

Often, researchers have combined multiple streams where each stream focus on different aspects of human activity recognition (Singh et al., 2016; Baradel et al., 2018b; Sharma et al., 2016; Aliakbarian et al., 2017). Ablation studies performed by the authors showed that by dedicating each stream for separate specialised tasks, these model were able to perform better than equivalent single stream architecture. Examples of multi-stream architectures include using separate streams for spatial and temporal processing (Sharma et al., 2016), separate streams for visual and flow features (Singh et al., 2016) and so on. Singh et al. (2016) used multi-stream architecture consisting of two CNN and one LSTM stream to model long temporal dynamics. The two CNN stream processed visual and flow features and the outputs of these two streams were passed to a LSTM which modelled the long-term temporal dependencies. The authors demonstrated that adding each stream to the

model positively impacted the overall accuracy. Similarly, Ma et al. (2016) proposed a three-stream network, where two of the streams focused on regions of interest while the third-stream concentrated on the optical flow. The novelty of this work (Ma et al., 2016) was in the integration of object-localisation in one of the visual streams which added to the model's scene understanding capability. Deng et al. (2016) combined an activity CNN with a person detection CNN to recognise group activities to present yet another example of multi-stream architecture. The authors (Deng et al., 2016) presented a novel graphical model that effectively combined a scene and a person stream. The graphical model helped to interpret higher-level scenes composition information in a semantic manner. Shi; Kim (2017) used a three-tier architecture composed of CNN and LSTM, which helped the model to combine depth and body-pose information effectively. The authors achieved superior performance through the use of privileged information which was a 'prior' added to the network through pre-training the input. Shi; Kim (2017) demonstrated that a purely data-driven model suffered from over-fitting and adding a 'prior' helped to mitigate the problem. It needs to be noted that this 'prior' or privileged information presents useful and enhanced data representation which is different from using a pre-trained model. The novel Spatial Encoding Unit (SEU) presented in the current study 'learns' an enhanced representation of the data, which consists of useful structural information about the body-pose. However, in contrast to Shi; Kim (2017), the SEU does not need pre-training and instead 'learns' the enhanced representation through integrated end-to-end training of the proposed model. Aliakbarian et al. (2017) built a multi-stage LSTM architecture that extracted features from ImageNet pre-trained VGG (Simonyan; Zisserman, 2014). The VGG features were fed into two different channels where one channel learnt action-aware information and other focused on context-aware features. An LSTM network combined the action-aware and context-aware features for early recognition of actions. The authors achieved early action recognition through the use of a novel loss function which encouraged high score for the correct class early. Instead of having multiple stream for handling or learning different tasks, specialised modules can be integrated within same stream to perform or learn special aspects of the data. Li et al. (2020) proposed a temporal excitation and aggregation block which included a motion excitation module and a multiple temporal aggregation module. The motion excitation module learns feature-level temporal differences from spatio-temporal features while temporal excitation and aggregation learns short temporal features and aggregates them to form a large receptive field. Both the modules complement each other to learn the overall temporal nature of the data.

Apart from multi-stream models, authors have also taken advantage of 'Attention' mechanism (Sec. 3.2.4) to improve performance of RGB video-based models. Baradel et al. (2018b) used 'Attention'-based interest points called 'Glimpse Clouds' involving ResNet-50. These 'Glimpses' consisted of important spatio-temporal points in a scene that are more relevant for discrimination. Sharma et al. (2016) proposed a model that integrated features from different parts of a spatio-temporal LSTM network and made 'Soft Attention'-based decision to recognise activities. The 'Attention' mechanism helped the model to selectively focus on parts of the video frames. The model essentially learnt which parts in the frames were relevant for discrimination and attached higher weightage to these parts. Inspired by these models, Self-Attention mechanism has been used in the current study (Chapter 6 and 7) to focus pre-trained CNN maps on important spatio-temporal points. Then, the Attention-focused maps are processed by LSTM to capture the temporal structures and dependencies

which provides better performance when compared to processing CNN maps directly (i.e., without 'Attention' mechanism).

### 3.4.3 Body pose-based models

Images or videos from monocular cameras do not contain depth information. With the availability of cheap depth sensors such as Microsoft Kinect (Han et al., 2013), depth information became readily available. As a result, 3D pose information extracted from depth-enabled devices like Kinect added another modality for activity recognition. Body pose data is normally available as 3D joint positions and is processed using temporal networks in the form of recurrent models such as LSTM (Shahroudy et al., 2016; Liu et al., 2016a). Liu et al. (2016a) improved the human tree structure model with the help of spatio-temporal features learned from a new gating mechanism of LSTM. The authors designed LSTM to be global context-aware by feeding contextual information in all steps. The model also selectively focused on more informative joints. Similarly, Song et al. (2017) introduced LSTM-based separate spatial and temporal 'Attention'-based networks for pose-based activity recognition. For each frame, the spatial network attached more weight to joints important to the current activity. Whereas the temporal network selected the more important frames. Shahroudy et al. (2016) introduced part-aware LSTM-based model for pose-based activity recognition. Instead of preserving memory of the entire body-pose sequence through LSTM, the model preserved groups of body joints representing body parts. By restricting the memory to 'learn' groups rather than the entire sequence the model was able to regularise the training better and hence performance better than traditional LSTM.

However, LSTMs suffer from few drawbacks which led authors (Lea et al., 2017; Kim; Reiter, 2017) to explore TCN-based pose sequence processing and analysis. In LSTM networks, individual cells are calculated for each time and take into account the output of the last time step only (Lea et al., 2017). This also means that each LSTM cell has to wait for the output of the last cell to process the information making LSTMs slow. As discussed earlier, TCN can be seen as 1D convolutional network combined with causal convolutions. Lea et al. (2017) first proposed TCN for activity recognition with an encoder-decoder and a dilated convolutions model. In contrast to LSTMs, the proposed 1D convolutions processed multiple time steps together making it faster and better equipped to capture long-range temporal dependencies (Lea et al., 2017). Kim; Reiter (2017) presented a TCN-ResNet model for action recognition with 3D skeleton sequences as input. The network incorporated residual connections with TCN, which has been widely used in CNN object classification models. The model has been evaluated on NTU 3D dataset (Shahroudy et al., 2016), which is currently one of the largest 3D action recognition datasets. Based on Kim; Reiter (2017), Xu et al. (2018) presented an ensemble of TCN-ResNets for skeleton-based human activity recognition. The final model was a score-fusion of multiple spatio-temporal and 'Attention'-based TCN-ResNets. In this study, TCN-ResNet (Kim; Reiter, 2017) has been used as the base network for the purely pose-based model presented in Chapter 8. The model adapts the TCN-ResNet to a two-stream spatial-temporal architecture. It combines the two-stream model with a FV-based intelligent pooling method to present an end-to-end trainable model that outperform the TCN-ResNet significantly. Similar to training models for object detection (Szegedy et al., 2015; He et al., 2016), pose estimation (Newell

et al., 2016; Cao et al., 2016a) and so on, authors (Demisse et al., 2018; Zanfir et al., 2013; Ke et al., 2017) have used data augmentation for increasing the recognition accuracy of pose-based models. Some of these are: augmenting coordinates with velocities and acceleration (Demisse et al., 2018; Zanfir et al., 2013), various normalisation techniques for the body joints (Zanfir et al., 2013) and adding relative positions (Ke et al., 2017). Ke et al. (2017) encoded the relative position of the joints in all the frames and passed the resulting vector into pre-trained CNN network. The authors (Ke et al., 2017) demonstrated that this enhanced representation improved the performance of their multi-task learning network. Inspired by the impact of enhanced representations, the current study proposes a novel SEU and a Temporal Encoding Unit (TEU). Instead of hand-crafting features for enhanced representation, the proposed SEU and TEU automatically 'learns' representations that can accurately capture the various inter-joint relationships and dependencies, and learn to recognise how these representations vary over time for various activity classes. Wang et al. (2012) split body pose information into five groups and then used spatial and temporal dictionaries to encode the spatial structure of human bodies. Vemulapalli et al. (2014) considered affine transformations to represent geometric relationships of body parts through Lie groups. The authors have extensively used RNNs for representing the skeleton sequences. Similarly, Du et al. (2015) used RNNs in a hierarchical manner to represent groups of body joints. Each joint has been represented by a sub-network at the initial layer; then the joint representations were fused hierarchically to form groups of joints. In a similar manner, Shahroudy et al. (2016) used body part-aware LSTM networks for encoding skeleton sequences. It is seen that authors have researched for better ways to encode the 3D joint positions for better spatial representation of the structural aspects of the human skeleton. The current study spatially encodes the joint positions through the SEU before feeding them into a 1D convolution-based network. This encoding is a sequence of augmented vectors that capture the structural representations between various body joints. Similarly, the proposed TEU augments the sequence in a temporal manner. Experiments demonstrate that the proposed novel SEU and TEU is able to impact the performance of the network positively.

More recently authors have used GNN (Kipf; Welling, 2016) for pose-based activity recognition. As the name suggests graph neural network is based on graphs. Yan et al. (2018) proposed a two-stream graph-based architecture. The first stream captured the spatial nature of the pose data while the second stream learnt the temporal nature. In spatial stream each joint was considered as a graph node and the bones connecting the joints were considered as vertices. In temporal stream each node represented the position of a joint in time and the vertices represented the temporal relationship of the joint. Authors have also combined video with pose information for human activity recognition (Baradel et al., 2018a). Video data offer important context cues, scene information, optical flow information and so on which can be used to enhance the performance of activity recognition models. However, given the large size of video data, it is difficult to achieve a meaningful combination. Out of the existing state-of-the-art approaches compared in Chapter 7 (Table 7.2 and 7.1) only Baradel et al. (2018a) and Shahroudy et al. (2017) used both body-pose and video data. In this work, a novel model is presented in Chapter 7 that successfully combines video and pose-based information. The ablation study (Chapter 7, Sec. 7.4.3) shows that the addition of a pose network to the video stream enhances the model performance.

## 3.5 Human activity recognition datasets

| Datasets | #Videos | #Classes | #Sub-classes | #Subjects | Data Modalities |
|---|---|---|---|---|---|
| MSRDailyActivity3D (Wang et al., 2012) | 320 | 16 | 0 | 10 | R,D,P |
| UTKinect (Xia et al., 2012a) | 200 | 10 | 0 | 10 | R,D,P |
| MSR-Action3D (Li et al., 2010) | 567 | 20 | 0 | 10 | R,D,P |
| CAD-60 (Sung et al., 2011) | 60 | 12 | 0 | 4 | R,D,P |
| CAD-120 (Koppula et al., 2013) | 120 | 20 | 0 | 4 | R,D,P |
| NTU-RGBD (Shahroudy et al., 2016) | 58K | 60 | 0 | 40 | R,D,P |
| Northwestern-UCLA (Wang et al., 2014a) | 1475 | 10 | 0 | 10 | R,D,P |
| Chardes (Sigurdsson et al., 2016) | 10K | 157 | 0 | 267 | R |
| NTU-RGBd 120 (Liu et al., 2019b) | 120K | 120 | 0 | 106 | R,D,P |
| Toyota Smart Home (Das et al., 2019) | 16K | 51 | 0 | 18 | R,D,P |
| UA-Concurrent Wei et al., 2020 | 201 | 35 | 0 | NA | R,D,P |

Table 3.1: Comparison of the proposed dataset with other activity recognition datasets. R: RGB; D: Depth; P: 3D Pose

Table 3.1 presents the well-known datasets popularly used to benchmark human activity recognition models. All the datasets present data in monocular RGB video, depth and human body-pose format. MSR daily activity (Wang et al., 2012) dataset was one of the earliest datasets to present Kinect-based body pose. As compared to more recent datasets such as the NTU-RGBD (Shahroudy et al., 2016), it is a small dataset with 320 sequences and 16 action classes. However, the author (Wang et al., 2012) proposed an evaluation protocol of using 160 samples for training and rest 160 for evaluation making a good case for generalisation. A good generalisation protocol and the small size makes it more suitable for preliminary training and evaluation. Thus, this dataset has been extensively used by many state-of-the-art approaches (Baradel et al., 2018a; Wang; Wu, 2013; Tao; Vidal, 2015; Zanfir et al., 2013). Therefore, in the current study, this MSR daily activity dataset (Wang et al., 2012) has been used to evaluate the proposed single-label activity recognition models against existing state-of-the-art approaches. On the other hand to prove a model's efficacy to the broader literature it is customary to use a large dataset (Baradel et al., 2018b; Baradel et al., 2018a). The NTU-RGBD dataset with around 56K sequences, which include approximately 16K samples for testing is one of the largest activity recognition datasets. Thus, in addition to the MSR-3D Daily Activity, this study also uses the NTU-RGBD (Shahroudy et al., 2016) dataset for the human activity recognition models. The main aim of this study is to recognise an ADL and differentiate between five different variations of the same ADL. The Table shows that the existing datasets only present distinctive activity classes whereas the current study requires a dataset that can present different intra-class variations of an ADL. This motivates and necessitates the multi-label activity recognition dataset presented in Chapter 5.

## 3.6 Conclusion

In this Chapter, the literature relevant to the proposed DL-based models in the current study has been discussed. It highlights that for spatial processing of images, authors have almost exclusively used CNN-based architectures. On the other hand, for processing temporal sequences authors have used mechanisms like LSTM and TCN. The wide use of CNN, LSTM and TCN provides the motivation for using these mechanisms in the current study. The discussion also highlights that in recent years 'Attention' mechanisms have been widely used to focus a network on more important points for discrimination thereby improving the model performance. The Chapter shows another aspect of DL-based model design called 'Pooling'. It explains how authors have focused on intelligent pooling for improving the performance of their activity recognition models. Inspired by this, the current study introduces a novel intelligent pooling method in Chapter 6. The discussion on activity recognition highlights the existing RGB video and pose-based approaches. For pose-based approaches, authors have used various encoding for enhancing the body-pose representation for better discrimination. This motivates the novel pose-encoding method presented in the current study (Chapter 7 and 8). The discussion on pose-estimation highlights the gap in literature with respect to lightweight human-pose estimation. This is the motivation to explore lightweight human-pose estimation in the next Chapter (Chapter 4).

# Chapter 4

# Lightweight Human Pose Estimation

## 4.1 Introduction

This Chapter attends to the second objective of this study which is to design a novel lightweight human pose estimation method. As seen from the literature review on CV-based physical rehabilitation methods, human pose estimation plays a vital role in CV-based rehabilitation and assessment. Patients undergoing physical rehabilitation require extensive monitoring and assessment over a period of time. This includes carrying out rehabilitation tasks at home which needs to be monitored and assessed over time. This has encouraged many researchers to develop methods for passive assessment which could be easily practised at homes (Natarajan et al., 2017). Many of these passive assessment methods make use of pose-based models that use Kinect as the primary sensor for obtaining human pose information. In this Chapter, the objective is to make use of the recent advances in DL to address this issue and provide an accurate assessment with a RGB sensor only. This makes it ideal for use in domestic environments, thereby providing an affordable healthcare alternative which is scalable. In the Chapter MobileNets (Howard et al., 2017), which is known to be a lightweight image classification network is adapted for mobile-based human pose estimation. Inspired by the widely used Stacked-Hourglass (Newell et al., 2016) type architecture for pose-estimation, MobileNets (Howard et al., 2017) is adapted for heat-map supervised training. Then, a novel 'Spilt-Stream' architecture is proposed at the final two layers of the MobileNets which reduces over-fitting and increases accuracy.

The next section elaborates the motivation further and illustrates the gap in literature that this Chapter aims to address. This is followed by a brief description of the MobileNets (Howard et al., 2017) and the Stacked-Hourglass network (Newell et al., 2016) which forms the basis of the current model. The subsequent sections describe the proposed approach in details (Sec 4.4), highlights the experiments conducted (Sec. 4.6) and demonstrate the results in comparison to existing state-of-the art approaches (Sec. 4.6). This is followed by an analysis of the 'Split-Stream' architecture which discusses the reason for its effectiveness (Sec. 4.8). The Chapter ends with a 'Discussion' section that connects the work to the broader literature.

### 4.1.1 Motivation/Rationale

Recent advances in CNNs have significantly influenced the performance of pose estimation models (Cao et al., 2017; Ning et al., 2017b; Papandreou et al., 2017a; Xiao; Wan, 2017). Most of these

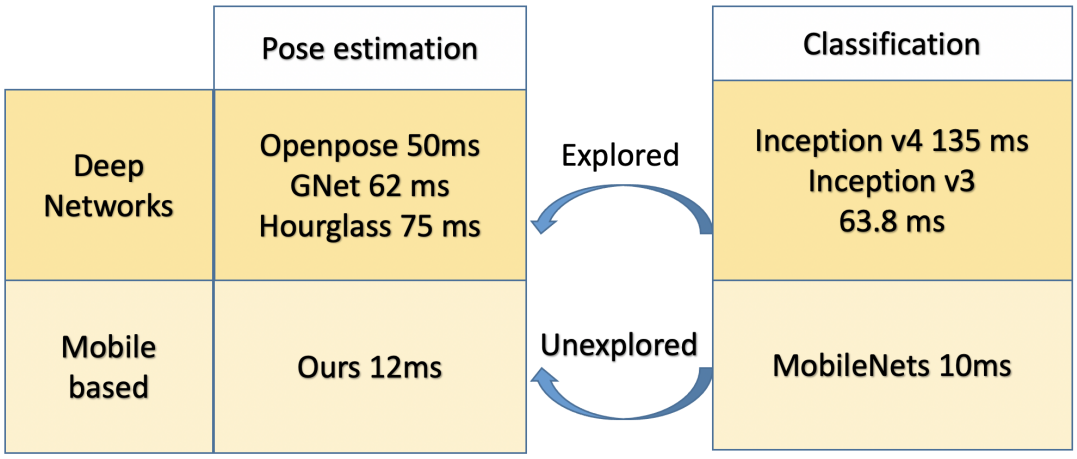



Figure 4.1: The study explores mobile-based pose estimation through adaptation of the lightweight MobileNets. Similar adaptations exist for larger models. GNet (Ning et al., 2017b) and Stacked-Hourglass (Newell et al., 2016) inference times are as reported in the paper. Inference times for Inception v3 (Szegedy et al., 2016), v4 (Szegedy et al., 2017) and OpenPose (Cao et al., 2017) are from the current setup.

models are complex and require powerful GPUs even for inference. In many real-world CV applications (home-based rehabilitation), there is a constraint on resources (e.g., power, memory) and the system is expected to work in real-time without compromising the accuracy. This trade-off is shown in Figure 4.1. As one can see, the area of lightweight mobile-based pose estimation is relatively less explored and the current study attempts to bridge this gap. High-performing DL methods for image classification (Szegedy et al., 2017) and pose estimation (Cao et al., 2016a) are known to have high inference time as they use large number of hidden layers having millions of tuneable parameters. These models also require high-performance GPUs for inference and are not suitable for mobile-based applications. For image classification tasks, many lightweight models (Redmon; Farhadi, 2017; Iandola et al., 2016; Chollet, 2017) have been developed for implementation on mobile devices. However, adapting these models for human pose estimation is still in its infancy.

One of the objectives in this work is to explore and adapt lightweight DL networks such as MobileNets for human pose estimation that can be easily deployed on mobile phones and other embedded platforms. The state-of-the-art models such as Inception-V3 (Szegedy et al., 2016) can achieve top-1 accuracy of 84% on Stanford Dogs (Khosla et al., 2011) dataset as compared to MobileNets' 83% (Howard et al., 2017). However, the number of parameters in MobileNets is $1/6^{th}$ of that of Inception-V3 (Howard et al., 2017). As shown in Figure 4.1, well-known pose estimation models take more than 50 ms for single image inference while MobileNets requires only 10 ms. Thus, this study intends to explore the area of mobile-based pose estimation by adapting MobileNets that have been used widely for classification tasks.

## 4.2 MobileNets Review

This section provides an overview of widely used MobileNets (Howard et al., 2017) architecture for developing lightweight DL models. All the arguments and equations in this section are referred from

Howard et al. (2017). There are two approaches for building small and efficient neural networks. The first approach is to compress pre-trained networks and the other is to train small networks directly. Most small networks are built for size but MobileNets is optimised for latency in addition to its small size. Speed is achieved by factorising the convolution operation. A form of factorising convolution operation called depth-wise separable convolutions was initially proposed by Sifre; Mallat (2014) and later adapted by Ioffe; Szegedy (2015) for Inception modules. The input feature map $F$ of a standard convolutional layer has dimensions $D_F \times D_F \times M$. which produces a feature map G parameterised by $D_F \times D_F \times N$. Here, $F$ is the width and height of feature map, $M$ is the input depth, $D$ is the width and height of a square output feature map and $N$ is the output depth. Let the convolution Kernel $K$ be parameterised by $D_K \times D_K \times M \times N$ where $D_K$ is the dimension of the square kernel. With an assumption of single stride the output feature map of a standard convolution operation can be formalised as:

$$G_{k,l,n} = \Sigma_{i,j,m} K_{i,j,m,n} \cdot F_{k+i-1,l+j-1,m} \tag{4.1}$$

The standard convolution has a cost of:

$$D_K \cdot D_K \cdot M \cdot N \cdot D_F \cdot D_F \tag{4.2}$$

The cost multiplicatively depends on the feature map $D_F \times D_F$, the kernel size $D_K \times D_K$, the number of input channels $M$ and the number of output channels $N$. MobileNets are based on a streamlined architecture that uses Depth-wise and Point-wise Separable Convolution (DPC). DPC breaks down the interactions in a standard convolution into smaller and computationally more efficient steps. In depth-wise convolution, a single filter is applied for each input channel. Then, a simple $1 \times 1$ convolution is applied to the output of each depth-wise convolution to create a linear combination of the output of the depth-wise layer. Formally depth-wise convolution can be expressed as:

$$\hat{G}_{k,l,m} = \Sigma_{i,j} \hat{K}_{i,j,m} \cdot F_{k+i-1,l+j-1,m} \tag{4.3}$$

Here $\hat{K}$ is the convolution kernel for depth-wise operation with size $D_K \times D_K \times M$. The $m^{th}$ filter in $\hat{K}$ is applied to the $m^{th}$ channel in $F$ to produce the $m^{th}$ channel of the filtered output feature map $\hat{G}$. The computational cost of depth-wise convolutions is given as:

$$D_K \cdot D_K \cdot M \cdot D_F \cdot D_F \tag{4.4}$$

DPC convolution computes the regular convolutional operation in two steps and the computational cost of the combined depth-wise and point wise convolutions steps is given as:

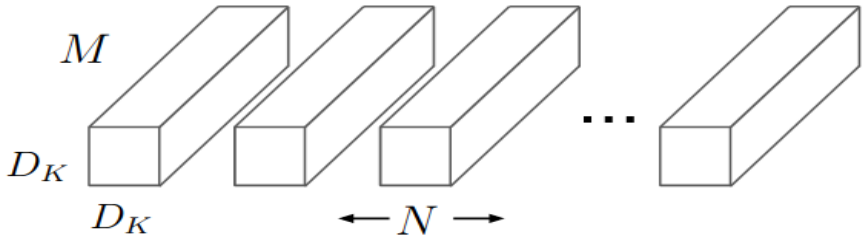
$$D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F \tag{4.5}$$

By expressing normal convolution operation as a two-step DPC, the authors achieve a reduction

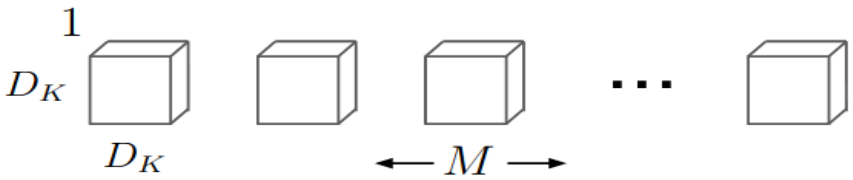
factor of:

$$\frac{D_K \cdot D_K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F}{D_K \cdot D_K \cdot M \cdot D_F \cdot D_F} = \frac{1}{N} + \frac{1}{D_K^2} \tag{4.6}$$

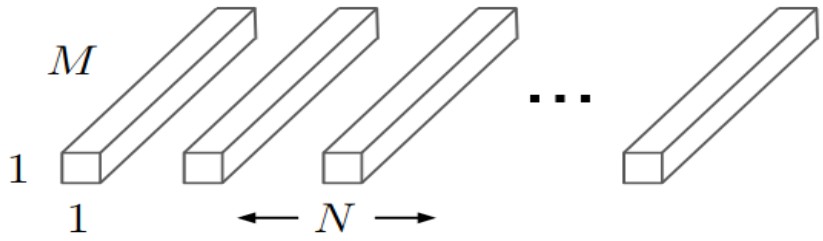
This results in drastically reduced numbers of parameters making the network faster. The network uses 3x3 depth-wise separable convolutions which result in up to 9 times fewer computations as compared to standard convolutions at the cost of a fractional reduction in accuracy. MobileNets lowers the resolution from $224 \times 224$ in the input layer to $7 \times 7$ in the last convolution layer but increases the number of filters from 32 in the first layer to 1024 in the penultimate layer. The last DPC layer is followed by a GAP layer whose output is reshaped and fed into a FC layer with an output size of 1000. The FC layer is responsible for 24.33% of the parameters.




(a) Standard Convolution Filters




(b) Depthwise Convolutional Filters




(c) $1 \times 1$ Convolutional Filters called Pointwise Convolution in the context of Depthwise Separable Convolution

Figure 4.2: The standard convolutional filters in (a) are replaced by two layers: depth-wise convolution in (b) and point-wise convolution in (c) to build a depth-wise separable filter. The Figure has been referred from Howard et al., 2017

To adapt MobileNets for pose estimation, the following factors were considered. Resolution of $7 \times 7$ at the final DPC layer makes heat-map regression difficult. Such supervision requires a higher resolution. But, the higher resolution with a large number of filters (1024) can be a speed bottleneck and may lead to over-fitting. The FC layer needs to be removed as it is unsuitable for heat-map regression. The intention was to implement these changes while still retaining part of ImageNet pre-trained MobileNets so that the model could benefit from the transfer learning. The GAP layer, which helps

to prevent over-fitting is not used in pose estimation architectures due to its incompatibility with heat-map regression (Cao et al., 2018; Newell et al., 2016). Thus, the goal was to introduce an alternate approach to prevent over-fitting.

## 4.3 Stacked Hourglass Network Review

As the name suggests, Stacked-Hourglass network (Newell et al., 2016) is a stack of hourglass modules. An hourglass module has an hourglass-like design where the resolution is gradually decreased and then increased. This is similar to an encoder-decoder architecture which is motivated by the need to capture information at every scale. The authors (Newell et al., 2016) suggest that whereas information is necessary to capture features like wrists, face and so on, a full coherent understanding of the whole image is required for final pose estimation. A person's body orientation, relative position of limbs and other such high-level information provides cues that are best understood at multiple scales. Figure 4.3 shows a single hourglass module of the Stacked-Hourglass network. The authors argue that any pose estimation network must have mechanisms to consolidate the information effectively across multiple scales. The need to capture information across scales is generally true for any image recognition and is followed by both classical methods such as SIFT (Lowe, 2004) or modern DL architectures such as Inception-ResNet-V2 (Szegedy et al., 2015), MobileNets (Howard et al., 2017). Thus, skip connections are introduced, which help in preserving spatial information at each layer. Skip connections are similar to residual connections which also serves the same purpose (He et al., 2016).

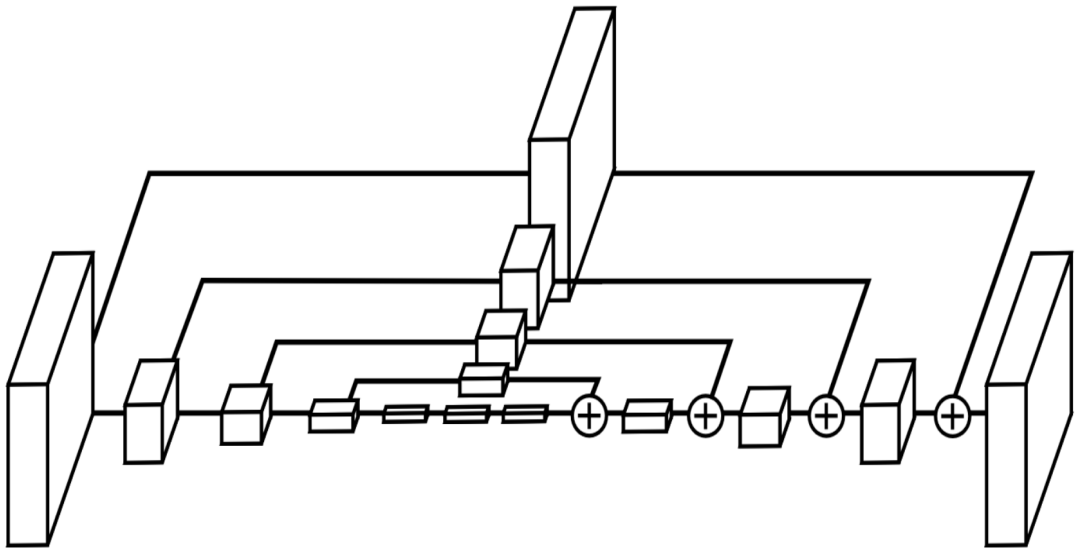


Figure 4.3: A single hourglass module of the Stacked-Hourglass network. The Figure has been referred from Newell et al. (2016)

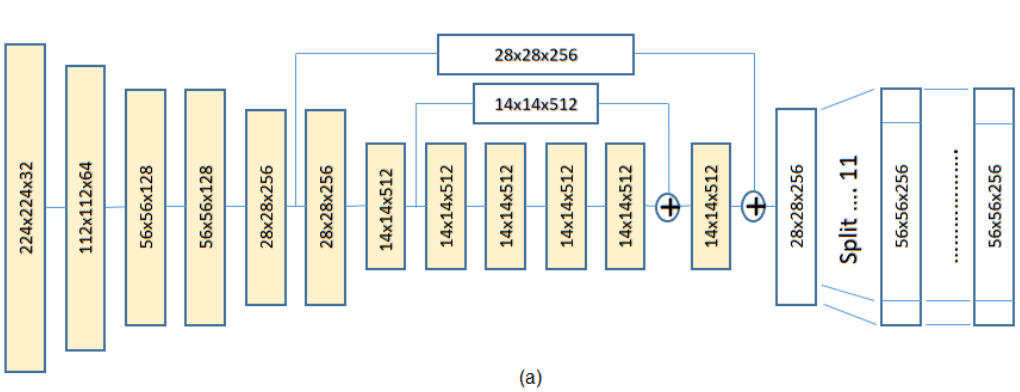
Now, the details of the hourglass design are discussed. A convolutional layer followed by a max-pooling layer is alternatively used to process input features down to a very low resolution. After each max-pooling step, the network branches off and applies more convolutions at the original pre-pooled resolution. Once the lowest resolution is reached, the network begins to up-sample the feature maps

and combine features across scales through the skip connections. Nearest neighbour techniques are used for the upsampling process. The skip connections are added in an element-wise addition manner to bring the same resolutions from the downsampling and the upsampling process, together. The topology of the hourglass is symmetric, which means for every layer present on the way down there is a corresponding layer going up. But for the model presented in this study, the input resolution goes from 256x256 to 7x7 and back to 65x65. Preliminary experiments suggested that increasing the resolution further makes the model heavy without any performance benefit.

## 4.4 Proposed Approach

To achieve the objectives, three mobile-based classification models, DenseNets (Huang et al., 2017), MobileNets (Howard et al., 2017) and SqueezeNets (Iandola et al., 2016) were considered. With 0.50 alpha and 160 × 160 input resolution, MobileNets has 1.32 million parameters (Howard et al., 2017) against SqueezeNet's 1.25 (Iandola et al., 2016). MobileNets scores 60.2% (Howard et al., 2017) while SqueezeNets scores 57.5% accuracy (Iandola et al., 2016) on ImageNet dataset. The best performing MobileNets variation with alpha 1 and input resolution 224 × 224 gives 70% accuracy (Howard et al., 2017). In this variation MobileNets has 4.2 million parameters (Howard et al., 2017). Thus, in this study MobileNets has been preferred over SqueezeNet due to higher accuracy. In experiments conducted for this project, DenseNets 121 took 63.8 ms while MobileNets took around 12 ms for the inference of a single image. DenseNets can be scaled from less than 1 to 20 million parameters. For a fair comparison size of DenseNets equivalent to that of MobileNets was chosen. A few different DenseNet block size combinations such as 6, 12, 12, 8 with 4.3 million parameters and 6, 12, 12, 16 with 5.7 million parameters were tested. Block size of 6, 12, 24, 16 was also tested with ImageNet initialised weights. Preliminary experiments showed that DenseNet did not converge as well as MobileNets.

### 4.4.1 MobileNets Modifications




| Joints | Filter distribution |
|---|---|
| Left shoulder | 25 |
| Left elbow | 27 |
| Left wrist | 35 |
| Right shoulder | 25 |
| Right elbow | 27 |
| Right wrist | 35 |
| Left hip | 27 |
| Right hip | 27 |
| Left eye | 10 |
| Right eye | 10 |
| Nose | 8 |

(b)

Figure 4.4: (a) Modified MobileNets architecture. First and last layers are normal convolution and rest are depth-wise and point-wise separable convolution blocks. Pre-trained lower layers from MobileNets are depicted in yellow. Last two layers are split joint-wise in the Split-Stream architecture. (b) Joint-wise filter distributions for last two layers

Inspired by the hourglass network (Newell et al., 2016), the final two DPC layers of MobileNets are

modified to increase the resolution through upsampling. If the whole model is changed to reflect an hourglass, the pre-trained weights cannot be used and the advantage of transfer learning would be lost impacting the accuracy. CNNs learn generalised features in the first few layers and class specific features towards the end. Thus, only the final two layers are changed where upsampling is used to increase the filter size from $14 \times 14$ to $28 \times 28$ and then to $56 \times 56$. Increasing the resolution further impacts the speed and thus, the final output resolution of the model is kept at $56 \times 56$. Lower size of heat-map resolution as compared to the input ($224 \times 224$) does not impact on accuracy as pointed out by Newell et al. (2016). To facilitate heat-map regression in the proposed model, the final filter resolution is $56 \times 56$ as opposed to $7 \times 7$ in the original MobileNets. To reduce the impact on speed due to increase in filter size, the number of filters are reduced in the final two layers. These layers have 256 filters each which is $1/4^{th}$ of the 1024 filters in the original MobileNets. For heat-map regression, the last FC layer is replaced by a normal convolution layer with 11 filters that correspond to heat-maps for 11 body joints. The two changed DPC layers with the final convolutional layer are shown in white towards the right of Figure 4.4a. The hourglass network also has side layers (skip connections) which are used to connect features across scales. The two horizontal white boxes depict the skip connections. These are introduced at resolution $28 \times 28$ and $14 \times 14$. Each skip connection goes through a DPC layer of the same dimension.

### 4.4.2 Split-Stream Architecture

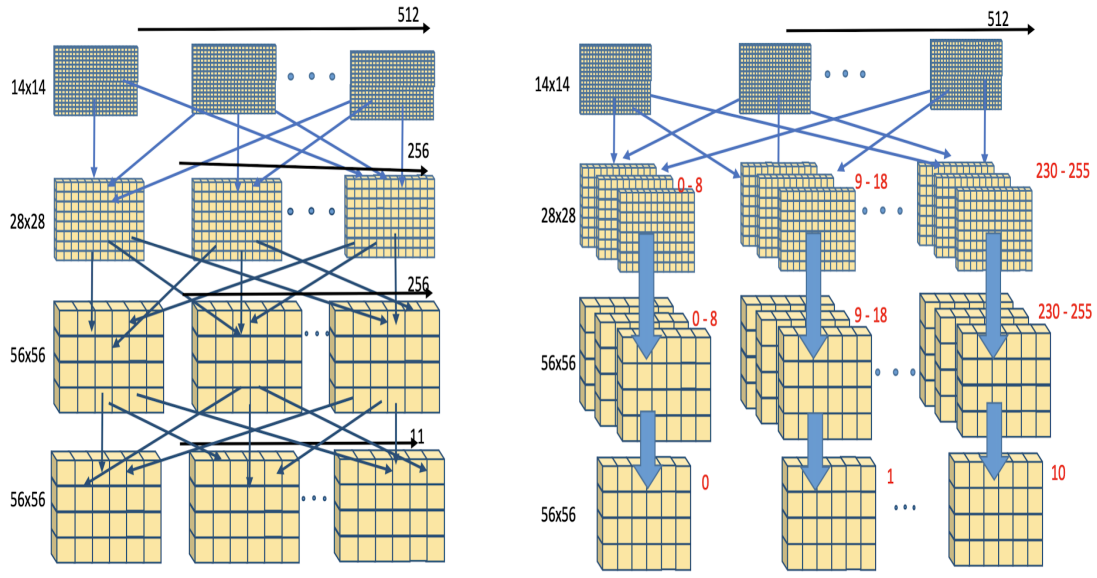



Figure 4.5: (a) Normal convolution operation in the last two layers vs (b) split-stream architecture. In split-stream architecture the convolution operation is split into 11 streams corresponding to the 11 joints.

GAP works by enforcing correspondence between confidence map and classes and it inherently prevents over-fitting (Lin et al., 2013a). GAP is not commonly used by pose estimation models (Yang et al., 2017; Ning et al., 2017b), since in regression problems there are no classes and thus this layer has been removed. The FC layer which is responsible for 24% of the parameters is already prone to over-fitting (Goodfellow et al., 2016), becomes more so in the absence of GAP and dropout. None of the standard pose estimation models (Newell et al., 2016; Yang et al., 2017) use dropout for dealing

with over-fitting as randomly dropping out filters is not suitable for regression. Instead, to deal with over-fitting a novel 'Split-Stream' method is introduced in this study. In this method, the last two DPC layers and the final convolution are split into 11 filter groups as shown towards the right side of Figure 4.4a. The same is also depicted in more details in Figure 4.5 where a comparison with normal convolution operation is shown. In normal convolution (Figure 4.5a) each of the 256 maps learn from all the 256 maps in the previous layer. Similarly, the final 11 heat maps corresponding to 11 joints learn from all the maps in the previous layer. However, in split stream architecture (Figure 4.5b), there are 11 streams. The 11 filter streams correspond to 11 upper body joints and the maps within any stream are shared but have no connection to the maps from different streams. As a result of the splitting, low-level features in the lower layers are common but high-level features of individual joints are regressed independently. This has two effects: 1) it reduces the number of parameters making the network lighter, 2) it reduces over-fitting for pose estimations problems where GAP or dropout is not used. The experiments 4.3 show that when 'Split-Stream' architecture is used the validation error follows the training error more closely than without it. To determine the number of filters needed for each joint, all the joints are first allocated filters equally. Then difficult joints like elbows and wrists are gradually allocated more filters than easier parts like the nose. Over several experiments the optimal filter numbers are obtained. The joint-wise filter distribution is shown in Figure 4.4b. In order to improve the detection performance of difficult joints, more filters were allocated to wrist and elbow joints but it did not increase accuracy any further.

## 4.5 Objective Function

Human pose estimation is a regression problem and thus the standard cross entropy error function (Goodfellow et al., 2016) is not applicable. Instead, a mean MSE has been used to regress the 11 heat-maps. Let $H$ indicate the ground truth heatmap and $\hat{H}$ produced by the model. There are $N = 11$ heatmaps for 11 joints. The size of each heatmap is $56 \times 56$. Therefore formally the loss function can be defined as:

$$L = \frac{1}{N} \sum_{i=1}^{N} (H_i - \hat{H}_i)^2 \tag{4.7}$$

## 4.6 Training Details

The well-known FLIC (Sapp; Taskar, 2013) dataset has been used for evaluating the proposed model. The dataset has been used by many well-known models 4.1 including the Stacked-Hourglass network (Newell et al., 2016) making it more suitable to compare the networks proposed network's performance. It consists of 5003 images out of which 3987 images are for training and 1016 are for testing. 80-20 split in a small dataset indicates good generalisation. Input images are cropped to loosely fit the person whose annotations are available and data augmentation is applied in the form of random rotation (+/- 30 degrees) and scaling (.75-1.25). For the baseline evaluation of MobileNets for pose estimation, only top Soft-max layer is removed and the number of classes changed to 22 ($2 \times 11$ body joints). MSE regression loss is applied to train the model. The performance of 'Split-

Stream' architecture is also evaluated with model supervised through MSE regression, although the final model is supervised with heat-map regression.

Figure 4.6: Example output from FLIC dataset. Predicted joint positions are marked in Red

Tensorflow is used for implementation along with Keras wrapper. For transfer learning supervision is carried out with the original layers frozen with a learning rate of 0.001 for 50K iterations, where only the new layers are trained which include the two layers receiving skip connections and split layers (marked in white Figure 4.4). Then the whole model is fine-tuned for 150K iterations, with the learning rate reduced to $1/10^{th}$. After the training loss plateaus, the learning rate was further reduced by half as is normally done. The standard practice is to use Stochastic Gradient Descent (SGD) for optimisation but Adam optimiser (Kingma; Ba, 2014) with default parameters was found to converge the model much faster. While optimising, the model also keeps track of moving averages

of the gradients with a decay of 0.9, which helps to stabilise the training by smoothing the changing of gradients. The model was trained on an Nvidia Quadro M4000 which has an effective memory of 6.7 GB, with a batch size of 16.

## 4.7 Evaluation

| Model | Elbows | Wrists |
|---|---|---|
| Toshev; Szegedy (2014) | 92.3 | 82.0 |
| Tompson et al. (2015) | 93.1 | 89.0 |
| Chen; Yuille (2014) | 95.3 | 92.4 |
| Wei et al. (2016) | 97.6 | 95.0 |
| **Proposed** | **97.6** | **95.2** |
| Newell et al. (2016) | 99.0 | 97.0 |

Table 4.1: FLIC results PCK@0.2

| MobileNets | Accuracy | Speed | Parameters | Size |
|---|---|---|---|---|
| Baseline | 96.4 | 10 ms | 4.3m | 68 MB |
| Split | 96.9 | 10 ms | 3.3m | 52 MB |
| Final | 97.3 | 12 ms | 2.3m | 26 MB |

Table 4.2: Comparison of proposed design with baseline

Evaluation is done using standard Percentage of Correct Keypoints (PCK) metric (Wei et al., 2016) where correct detection falls within 20% of torso size from the ground truth. For comparison wrists and elbow detection rate are reported as these are the most difficult joints and are widely used for the performance comparison on FLIC dataset. Table 4.1 compares the results with other models and shows competitive results although our model is optimised for both speed and accuracy rather than only accuracy. State of the art result achieved by the Stacked-Hourglass network (Newell et al., 2016) takes 75 ms for a forward pass on a 12GB Nvidia Titan. whereas the proposed model takes 12 ms for a forward pass on 8GB Nvidia Quadro. Moreover, the real advantage of using MobileNets based model is that it is optimised for mobile-based CV applications. The FLIC dataset used for training the model is very basic containing single person upper body joints only. Recent pose estimation models such as Alpha pose (Fang et al., 2017), have used advanced multi-person datasets such as COCO (COCO, 2016). Results presented in Table 4.2 have not compared recent more methods as recent models have not used the FLIC dataset.

Table 4.2 compares the baseline performance with the 'Split-Stream' architecture and the final proposed model, which is the novel design that combines the 'Split-Stream' architecture and the hourglass network. The baseline model is transfer learned from ImageNet pre-trained MobileNets (Howard et al., 2017) and performed at an accuracy of 96.4% on the FLIC dataset. If trained from randomly initialised weights the performance is much lower (~87%). With the application of the 'Split-Stream' architecture but still regressing with MSE, the gain in accuracy is 0.5%. The final accuracy gain with 'Split-Stream' architecture and hourglass-inspired design is 0.9%. The number of parameters

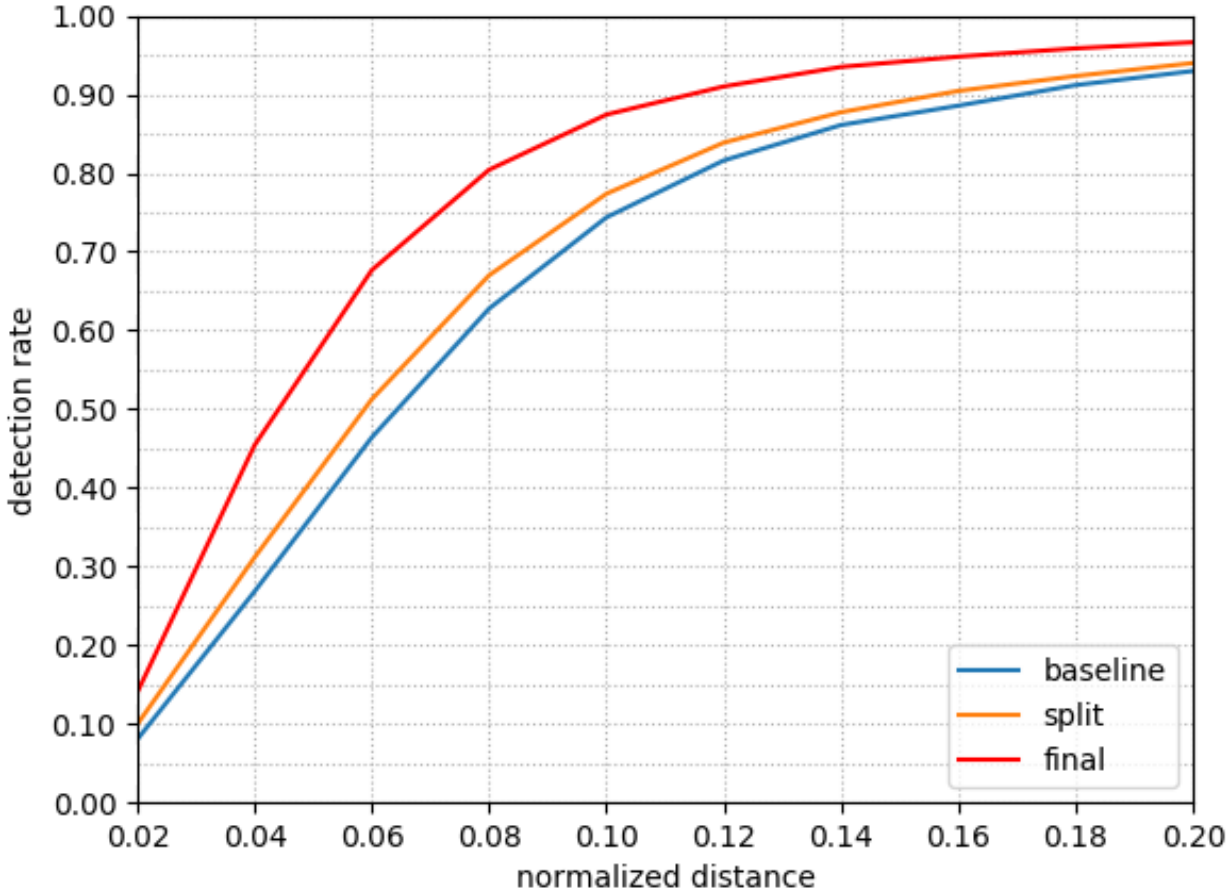



Figure 4.7: Comparison of elbow and wrist accuracy with baseline across PCK thresholds. Baseline: Regression on MobileNets (Howard et al., 2017). Split: Introduction of the split-stream architecture in the final two layers. Final: Modification of MobileNets to represent Hourglass (Newell et al., 2016) network with heat-map regression

and parameter size of the proposed model is approximately half of MobileNets. The marginal drop in speed is mainly due to the heat-map regression. Figure 4.7 shows the proposed method performing much better than the baseline at lower PCK thresholds.

## 4.8 Split-Stream Architecture Analysis

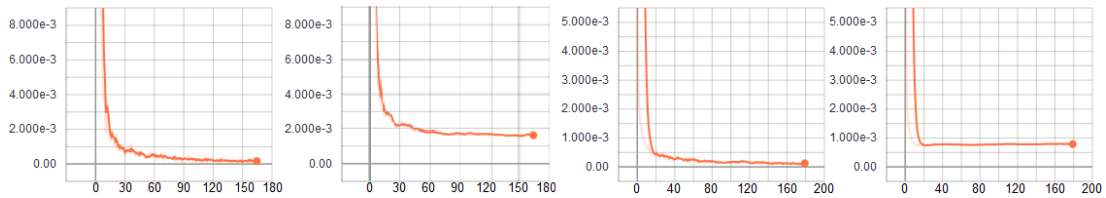



Figure 4.8: Loss (Y-axis) vs iteration in 1000s (X-axis) curve as generated by Tensorboard. From left to right: MobileNets train loss; MobilNets validation loss; proposed model train loss; proposed model validation loss.

The main novelty of the proposed model lies in the 'Split-Stream' architecture and this section further elaborates why it works. The network is split into separate groups of filters that do not

share weights with other filter groups. This implies that the filter groups at the final two layers take common low-level representations of the whole image as input but learn each joint independently of other joints. Table 4.2 shows that this brings down the number of parameters and parameter size by approximately half, which is a big advantage for mobile-based applications as the memory requirement shrinks with reduced parameters. Reducing over-fitting is another advantage of this 'Split-Stream' design. It is a well-known fact that larger networks are prone to over-fitting (Szegedy et al., 2017). This ultimately led to the formation of inception modules (Szegedy et al., 2015), which uses reduced connections. Figure 4.8 compares the train and validation error of MobileNets and the proposed design. Even though the final training loss is the same for both cases, the validation loss for the proposed model is less than half of the original MobileNets. MobileNets uses GAP for preventing over-fitting. When the GAP layer is removed, the validation error does not change much but the training error drops by a factor of 6 along with 0.5% reduction in accuracy as shown in the second row of the Table 4.3. This is a typical sign of over-fitting (Goodfellow et al., 2016). When the last two layers are split into separate filter groups for each joint, both accuracy and validation error improves and the model performs better than the baseline. This shows that the 'Split-Stream' design helps in reducing over-fitting. It is interesting to note that the accuracy of each joint is only loosely tied to number of filters allocated. Stacked-Hourglass

| MobileNets | Train error | Val error | Accuracy |
|---|---|---|---|
| Baseline | 1.57e-4 | 1.67e-3 | 96.4 |
| GAP removed | 2.4e-05 | 1.04e-3 | 95.9 |
| Split | 1.52e-4 | 7.929e-4 | 96.9 |

Table 4.3: Modified MobileNets (Howard et al., 2017) comparison. Baseline

## 4.9 Discussion

In this Chapter, the area of lightweight human pose estimation has been explored. The proposed model demonstrates the adaptation of well-known fast and efficient MobileNets for human pose estimation through transfer learning. The network is adapted for heat-map regression inspired by the Stacked-Hourglass network (Newell et al., 2016). It also introduces a novel 'Split-Stream' architecture, which reduces over-fitting (Sec: 4.8) and could be a potential alternative to the customary GAP layer present towards the end of many CNN models (Howard et al., 2017; Szegedy et al., 2015; Szegedy et al., 2016; Huang et al., 2017). The proposed model outperforms the baseline with GAP considerably across PCK thresholds (Figure 4.7). As mentioned in the section (Sec: 4.1), the area of lightweight object detection/classification (Howard et al., 2017; Huang et al., 2017; Iandola et al., 2016) has been extensively explored by researchers but the area of lightweight pose estimation is yet to be fully explored. The experiments (Table 4.1) demonstrate that the proposed model achieves close to state-of-the-art results while having fewer parameters and inference time (Table 4.2) and therefore more efficient.

In relation to the project, this Chapter addresses the second objective, which is to present a lightweight

mobile-based human pose estimation model. Patients undergoing physical rehabilitation require passive home-based monitoring, of which human pose estimation is a critical part. It is not feasible to have high-performing GPUs in such home-based or mobile devices due to cost and size issues. In such cases mobile-based pose estimation is highly desirable. Thus, the proposed model will help advance the field of mobile and embedded CV applications focusing on human pose estimation.

## 4.10 Conclusion

The proposed model demonstrates the adaptation of well-known fast and efficient MobileNets for human pose estimation through transfer learning. The main contribution of the proposed model is a novel 'Split-Stream' architecture which helps in reducing over-fitting. The lightweight pose-estimation model has the potential of application in home-based rehabilitation for physically impaired persons. The MobileNets based pose estimation model presented in this Chapter has been presented and published in the conference proceedings of the *15th IEEE International Conference on AVSS, 2018*.

# Chapter 5

# Functional Activity Recognition Dataset

## 5.1 Introduction

This Chapter is targeted towards the third objective of this research, which is to present a multi-label ADL recognition dataset that presents physical impairment-specific versions of ADL. The main aim of the research is to develop an AI or DL model that can recognise an ADL and discriminate between the regular and various physical impairment-specific versions of the same ADL. In the previous chapters (Chapter 1, Sec. 1.1.2, Chapter 2 Sec. 2.9, Chapter 3 Sec. 3.5) the importance of large-scale publicly available benchmark datasets in developing DL models has been highlighted. The discussions also highlight that such datasets are not readily available for functional assessment through ADL.This makes it difficult to develop models that can be used to improve automated assessment of physically impaired persons. Thus, the third objectives of this work is to fill this gap by contributing a novel dataset that can be used for this purpose. The multi-modal dataset presented here contains 5685 samples of 10 common ADL presented in RGB video, depth and human-body pose format. For each ADL, the dataset presents the normal and four different physical impairment-specific versions. Thus, each sample presented has two labels, one for the 'Activity' (e.g., drinking, walking) and the other one for the 'Impairment' (e.g., normal, ataxic) and hence the name multi-label ADL recognition dataset. The rest of this Chapter is organised as follows. The next section elaborates the rationale behind the formulation of this dataset. The dataset design including specification and constraints are described in section 5.2. Then, each of the 'Activities' and the impairments and their formulation for the filming is discussed. The subsequent sections discuss the data collection methodology, cleanup, post-processing and technical details that users will need to understand to use the dataset.

### 5.1.1 Motivation/Rationale

Researchers have approached automated assessment in various different ways which have been described in Chapter 3. To summarise, some of the common approaches to assess a patient's condition include comparison of normal and abnormal joint angle trajectories (Chapter 2, Table 2.4). Other researchers have used gesture recognition to understand if a patient can attain certain therapeutic postures (Chapter 2, Table 2.5). Authors have also formulated the problem as activity recognition to ascertain a patient's condition (Chapter 2, Table 2.4). However, Chapter 2 (Sec. 2.11) points that researchers are yet to explore automated functional assessment of a physically impaired person through ADL. This study aims to improve automated functional assessment of ADL, by recognising

an ADL and discriminating between a normal and different physical impairment-specific versions of the same ADL. Chapter 2 also shows that most researchers have trained and evaluated their model on small datasets specifically designed to fit their approach. In contrast to other areas of CV (e.g., human pose estimation, human activity recognition) their are a very few publicly available datasets and these have been tabulated in Table 2.7. The table shows that these datasets are targeted towards specific body movements such as a knee, UPDRS compensatory movements and so on. These datasets are not intended for functional assessment of patients through ADL and therefore do not present ADL, which are a more generalised form of human activity. ADL recognition has been extensively explored by the CV community and Chapter 3 presents the relevant literature (Sec. 3.4) and existing datasets (Sec. 3.5). Existing research and datasets present ADL as performed by healthy persons and do not take into account their physical impairment-specific variations. Physically impaired persons would perform an ADL differently from healthy persons and thus would present the same ADL in a different manner depending on the type of impairment. For example, a person having tremors would shake his or her hand while drinking water whose spatio-temporal trajectory would be different from a drinking action without tremors. The existing datasets (Table 2.7 (Chapter 3.4 ) are not appropriate to validate solutions developed to address this issue. Thus, this study presents a dataset that captures the difference between the normal and various physical impairment-specific versions of the same ADL.

## 5.2 Dataset Design

Human motion manifests in a wide variety of forms and so does its abnormalities. Due to such varying manifestations of human movement and abnormalities, it is not feasible to capture the whole range of ADL and their corresponding impairments. The idea was to prepare a dataset that meets the following constraints:

- The dataset should contain enough 'Activities' that would collectively cover a wide range of body movements and capture a few common abnormalities.
- The dataset should contain enough samples that would suffice the needs of developing AI-based models.

**Constraint 1:** To assess a patient's condition and to determine their functional independence, clinicians often require them to perform day-to-day activities or ADL (Edemekong PF, 2020; Green; Young, 2001). The initial idea was to capture patients while performing these ADL and label each action with an 'Activity' and an 'Impairment'. For example, if a patient performs the act of drinking water with tremors then the sequence could be labelled as 'Activity = drinking' and 'Impairment = tremors'. There were two obstacles to the proposed idea: 1) Ethical clearance and time constraint; 2) Uniformity of sequences required for a dataset. It is reasonable to assume that within the short timeframe of the research it would be infeasible to obtain ethical clearance and consent of patients to film them, as clinicians are assessing them. Second, to create a dataset one needs multiple samples of the same label, ideally across a number of subjects. For example, if the requirement is to create a dataset

with two 'Activities' (sit to stand and walking) with two 'Impairments' (normal, bent-knee and wider gait), then the dataset ideally should have an equal number of repetitions for each 'Activity' and 'Impairment' combinations across a number of subjects. Again, it would be impractical to ask patients to perform multiple repetitions of each of these ADL owing to their physical constraints. It is easy to see that a patient with a bent-knee, would face difficulty in performing sit to stand multiple times and would be unable to provide a regular sit to stand sample. The workaround was to film the ADL with healthy subjects while acting like patients. To make sure that ADL performed by healthy subjects accurately reflects performance of real patients, help was sought from an Occupational Therapist. Dr Helen Carey, who is a professional lead in Occupational Therapy at the Wrexham Glyndwr University, kindly agreed to guide the participants. Under her guidance, common 'Impairment's were identified that patients exhibit while performing ADL. Dr. Carey provided video samples of how actual patients would perform ADL with these impairments. Under her guidance, 10 different ADL including a normally executed and four different impairment-specific versions of each ADL were selected for the dataset. The ADL were chosen in a manner that would collectively cover a wide range of body movements and test various parts of the musculo-skeletal system. Table 5.2 lists the ADL along with the impairments. 'Sitting', 'Standing' and 'Walking' cover lower torso and leg movement, while the other ADL test a subject's ability to move their upper limbs and upper torso. 'Brushing Floor', 'Answering Phone' and 'Clapping' are performed while standing and thus they require close co-ordination between upper and lower halves of the body.

**Constraint 2:** The goal was to design a dataset that would be feasible to capture within the time frame of the project and would suffice the needs of today's AI-based models. Three factors: i) overall sample size ii) number of classes and iii) number of participants were considered to determine the size of the dataset. Table 3.1 (Chapter 3, Sec. 3.4) shows a list of well-known publicly available datasets for human activity recognition. The MSR 3D Dataset (Wang et al., 2012) containing 320 videos and 16 classes is considered a small dataset. In contrast, the NTU RGBD (Shahroudy et al., 2016) dataset consisting of around 56K samples and 60 different classes is considered as a large dataset. Considering the above factors, the study aimed to collect a dataset of around 5K samples. The goal was to film the 5K samples with 10 different subjects for 10 different ADL with 5 variations of each ADL. Thus, the study aimed to capture 50 different 'Activity-Impairment' combinations which is significantly more than most of the datasets in Table 3.1 (Chapter 3, Sec. 3.4). The methodology section 5.4 discusses the dataset dimensions in more details. As shown in Table 2.7 (Chapter 2 Sec. 2.9) and Table 3.1 (Chapter 3 Sec. 3.5) most of the existing dataset have filmed the data in a multi-modal format including RGB, depth and 3D human body-pose. Thus, it was planned to film the current dataset with Kinect which provides data in the above-mentioned formats.

## 5.3 Impairments and Activities

This section first describes each impairment, including its underlying cause(s) and its formulation in the context of this dataset. The second part describes each ADL and the body parts involved in each of them. In addition to that, the executions of impaired versions of the ADL is also duly described.

### 5.3.1 Impairments

Figure 5.1: Ataxia: Clockwise from top left, sequence of snapshots shows 'Ataxic' walking with arms swinging and torso rotating to imitate involuntary movements

**Ataxia:** Ataxia is an umbrella term for a group of disorders that can affect physical co-ordination, balance, speech, vision, swallowing and so on (NHS, 2018). Damage to a part of the brain called the cerebellum can cause Ataxia. This may be due to an underlying condition such as Multiple Sclerosis or may be due to an injury. In most cases, there is no treatment for Ataxia but physiotherapy may help with movement related issues. Although in Ataxia any part of the body may be affected, for this dataset, the main concern is balance and co-ordination while performing ADL. Subjects display involuntary movements and lack of balance while performing ADL. As shown by the Occupational Therapist, the subjects imitated involuntary sideways and rotatory movement of the torso and uncontrolled limb movements. This impairment is included for all the 10 ADL.

**Elbow Rigidity:** A rigid or 'stiff' elbow may be as a result of injury or other conditions like tendonitis (for example tennis elbow), sprains and strains or arthritis (NHS, 2017). This may also occur due to conditions like frozen or inflamed (bursitis) shoulders, although the frozen shoulder aspect is not considered for this dataset. As the name suggests, patients exhibiting this condition are

Figure 5.2: Elbow Rigidity: Snapshots from 'Answering Phone', illustrate that the activity is completed with little or no elbow flexion or extension

unable to extend or flex their elbow. Patients exhibiting such a condition often exhibit compensatory movements to perform an ADL or achieve other functional goals (Lee, 2015). When patients move body parts other than the affected part to achieve a functional goal which would not be normally needed, the movement is called 'compensatory movement'. For example, a patient may move the neck forward to drink water with bent-elbow, to meet the water bottle with his or her mouth. For 'Elbow Rigidity' and 'Knee Rigidity' the dataset captured such movements as shown by the Occupational Therapist. In this dataset, subjects imitate what is commonly known as bent-elbow where, a subject's elbow is slightly bent to begin with and cannot flex or extend their elbow. This impairment is captured for all the upper limb ADL as shown in Table 5.2. Subjects also exhibit the compensatory movements (Lee, 2015) as described in activity definitions that follow.

Figure 5.3: Knee Rigidity: From left to right 'Standing', 'Walking' and 'Sitting'. Subjects imitate bent-knee wherein the knee is rigid in a bent position. Compensatory movement is provided by raising ankle and the majority of the body-weight is carried on the other leg

**Knee Rigidity:** Similar to 'Elbow Rigidity', 'Knee Rigidity' is a term for the condition that occurs when a patient is experiencing stiffness in the knee and is unable to flex or extend the knee. This can be either due to injury or due to conditions like osteoarthritis, weak muscles, overuse and so one. Subjects imitate what is commonly known as 'bent-knee' where the knee is stiff in a bent position and is neither flexible or extensible. In this dataset 'Knee rigidity' is exhibited for the three ADL involving the lower limbs as shown in Table 5.2. Similar to 'Elbow rigidity' various compensatory movements are captured and are described in the respective activity descriptions.

**Tremors:** Tremors are uncontrollable and involuntary shaking of the hands or other parts of the body. Tremors may occur due to age related factors and may not require medical attention. Conditions like Parkinson's Disease, Overactive Thyroid, Multiple Sclerosis, Dystonia, Stroke and Peripheral Neuropathy can also cause tremors which may require medical attention (NINDS, 2020). There is generally no cure for tremors, although medications and physiotherapy help to manage symptoms. In this dataset, 'Action tremors' (NINDS, 2020) is exhibited for all the upper-limb ADL as shown in Table 5.2. Action tremors occur with voluntary movements of muscle, for example while performing an ADL.

Figure 5.4: Wider Gait: From left to right 'Wider Gait' vs 'Normal' stance while 'Walking two steps'

**Wider Gait:** Wide-based gait, broad-based gait or 'Wider-Gait' is a type of gait abnormality in which the feet are wider apart while walking than normal. This type of gait abnormality commonly associated with elderly patients (Pirker; Katzenschlager, 2017). Elderly individuals have a 40% 'Wider-Gait' than younger persons. Ataxia, other cerebellar diseases and Myelopathy can also cause Wider Gait. In addition to positioning their feet wider apart, the movements look clumsy and unstable. For the dataset, this impairment is present in the three lower limb ADL, as shown in Figure 5.2.

**Shoulder Weakness:** Weakness in the shoulder is a very common condition where one or both the shoulders become too weak to fully support the range of shoulder movements required to carry out functional ADL. Old age, injury, shoulder impingement, nerve damage are some of the causes responsible for shoulder weakness. In this dataset, subjects imitate scenarios where one shoulder is too weak to carry out shoulder elevation movements. As a result, subjects tend to lean towards the weak shoulder side while performing ADL with the arm involving the other (normal) shoulder. The same is illustrated through 'Clapping' activity in Figure 5.5. In this dataset, this impairment is included in all the hand ADL.

Figure 5.5: Clapping: From left to right 'Weak Shoulder' vs 'Normal' stance while 'Clapping'. To imitate 'Weak Shoulder' subject leans towards the weak shoulder side and do not lift the arm as well as the normal arm

Figure 5.6: Weakness to one side: From left to right 'Walking', 'Standing' and 'Sitting'. Subjects lean towards the weaker side while displaying very little movement on that side

**Weakness to one side:** Medically known as Hemiparesis, this impairment is caused when muscles in one side of the body become partially weak. The weakness may involve muscles in the leg, arm,

face or any combination of these. Typically, injury to the left side of the brain causes Hemiparesis on the right side of the body and vice versa. Hemiparesis is different from Hemiplegia where one side of the body is completely paralysed (NCBI, 2020). Stroke is the most common cause of Hemiparesis and 80% of stroke survivors experience Hemiparesis (ASA, 2019). Apart from stroke, conditions like a tumour, Multiple Sclerosis and traumatic injury and so can cause 'Weakness to one side'. This impairment is included in all the lower limb ADL in this dataset. To imitate 'Weakness to one side' subjects lean towards the weaker side and keep the movements in the weaker side as small as possible.

### 5.3.2 Activities

**Walking two steps:** In this activity, subjects start from a standing position with their feet together. Moving one step at a time the subject takes two steps and end with their feet together. This ADL mainly tests the movement of the lower portion of the body. The dataset presents the following impairments for this activity:

- **Ataxic:** For this impairment, subjects imitate involuntary shaking and movement for the whole body while taking a step and the shaking continues even after the step is completed.
- **Knee Rigidity:** Subjects imitate a rigid, slightly bent-knee on the right leg. The ankle is raised as compensation to support the bent-knee and most of the load bearing is done through the normal leg.
- **Weakness to One Side:** To imitate this impairment, subjects lean towards the weaker side and droop their shoulder while making as little movement as possible on the weaker side.
- **Wider Gait:** In addition to keeping the feet wider apart subjects tend to sway from side to side and the movements look a little clumsy.

**Sitting:** For this activity subjects sit on a chair without leaning to the backrest and attempt to stand up without arm support. This activity also mainly tests the lower body functional independence. The impairments involved are the same as for 'Walking two steps'.

- **Ataxic:** For this impairment, subjects imitate involuntary shaking and movement for the whole body as they try to sit down. The shaking continues after the action is complete.
- **Knee Rigidity:** Here also, subjects imitate a rigid, slightly bent-knee on the right leg. The ankle is raised as compensation to support the bent-knee and most of the body weight is supported by the other leg during the sitting actions. The subjects tend to fall back slightly as they sit since there is little or no support from the leg having the bent-knee.
- **Weakness to One Side:** To imitate this impairment, subjects lean towards the weaker side and droop their shoulder and sometimes take the support of the chair with their hand while performing the sitting action.

- **Wider Gait:** In addition to keeping the feet wider apart subjects tend to sway from side to side and the movements look clumsy. Subjects take support with their hands on both knees to stand up.

**Standing:** In this ADL, the subject goes from a standing position to sitting on a chair. This is the third and final activity which mainly tests the lower body functional independence. The impairments involved are the same as 'Walking two steps' and 'Standing'.

- **Ataxic:** In a manner similar to sitting, subjects start shaking and display involuntary movements as they attempt to perform the activity and continue even after the activity is over.
- **Knee Rigidity:** Similar to 'Walking two steps' and 'Sitting', subjects imitate a rigid and slightly bent-knee on the right leg. To start with, the ankle related to the affected knee is slightly raised and it straightens as the subjects stand up, to support the bent-knee. Most of the load bearing is carried out by the normal leg.
- **Weakness to One Side:** Just like sitting, subjects lean towards the weaker side, droop their shoulder and sometimes take support of the chair while performing the standing action.
- **Wider Gait:** Similar to sitting, subjects keep their feet wider apart, tend to sway from side to side and the movements look a little clumsy while performing the activity. In addition to that the arms take the support of the knees to help the subject stand.

**Clapping:** This activity tests the functional independence of upper limbs and tests a subject's ability to raise their arm. Subjects are required to stand in an arms down position and then raise their arms to clap twice.

- **Ataxic:** For this impairment, subjects shake their arms in addition to the body (torso) while performing the activity.
- **Elbow Rigidity:** For this impairment subjects imitate rigid and slightly bent-elbows. Subject raise their arms bent at the elbows and perform the clapping action exclusively with shoulder movements.
- **Shoulder Weakness:** For this impairment subjects lift the unaffected arm and bend the affected arm from the elbow without lifting the shoulder. This results in a vertical clap as opposed to a normal horizontal clap. This imitates a 'Weak Shoulder' scenario where patients are unable to perform the shoulder elevation movement.
- **Tremors:** As the arms go up to perform the clapping activity, both the hands begin to shake with the movements more pronounced towards the palm. Unlike Ataxia, the rest of the body is stable.

**Reaching above:** This is the second activity that tests functional independence of the upper limbs while testing the ability of a subject to raise their arm above their head level properly. Subjects perform the functional act of reaching above with their arms to clean or reach for objects.

- **Ataxic:** Like other activities, subjects shake their whole body to display involuntary movements. In addition to that, the hand in action also keeps shaking which makes reaching above to the required extent difficult.
- **Elbow Rigidity:** Subjects start with a bent-elbow and raise their affected arm to reach above. It is difficult to reach the desired height when the arm is bent at the elbow and subjects raise their heels to compensate for lack of reach.
- **Shoulder Weakness:** For this impairment subjects lift the unaffected arm to reach above and in the process lean towards the weaker shoulder side. The weaker shoulder is unable to support the body reaching above with shoulders parallel to the ground causing the body to tilt.
- **Tremors:** As the arms go up to perform the clapping activity, both of them begin to shake with the movements more pronounced towards the palm. Unlike Ataxia, the rest of the body is stable.

**Answering phone:** Along with 'Drinking', 'Brushing hair' and 'Wearing Glasses', 'Answering Phone', investigates a subject's ability to move a hand in coordination with neck movements. To answer a phone call, the subjects reach out to pick a phone placed on a chair by their side and lift it near to their face.

- **Ataxic:** Like other activities, subjects shake their whole body to display involuntary movements. In this activity, subjects face difficulty while picking up the phone and then keep it stationary close to their ear.
- **Elbow Rigidity:** It is very difficult to perform this activity with a rigid, bent-elbow. The subjects use trunk and/or knee compensation for lifting the phone as they face difficulty in reaching the phone. To reach the ear, subjects cannot flex the elbow to the required extent and compensate by tilting their neck towards the phone as shown in Figure 5.5.
- **Shoulder Weakness:** Subjects performing this activity with the normal hand are slightly tilted towards the side having the weak shoulder. This is because the weaker shoulder is unable to support the body's normal movement pattern and droops causing the neck and body to lean towards it.
- **Tremors:** As in other activities the hand keeps shaking while the subject is answering the phone call. Subjects face difficulty in picking up the phone and holding it steady near their ear.

**Brushing Hair:** Similar to 'Answering phone' this activity requires co-ordination of hand movement with the neck/head. Subjects perform this activity in a sitting position with their hands resting on their hips. The right hand is then lifted to reach the hair and a single brushing stroke is performed, after which the hand comes back down to the starting position.

- **Ataxic:** Subjects imitate involuntary movement throughout the body as well as the hand performing the activity. This makes it difficult to do a proper brushing stroke, which is what an actual patient with ataxia will face while brushing their hair.

- **Elbow Rigidity:** Subjects start with a bent-elbow and lift their affected arm to reach for the head. Subjects face difficulty in reaching the head and compensate by tilting the neck towards the hand. The brushing stroke is mainly performed by shoulder movement.
- **Shoulder Weakness:** Similar to other activities subjects perform this activity with the unaffected hand while the other shoulder droops and the neck tilts towards the weaker shoulder.
- **Tremors:** As the hand reaches the head while shaking, subjects are not able to execute a proper brushing stroke and struggle with brushing the hair properly

**Drinking:** This activity is performed in a sitting position with the subject holding the bottle with their right hand to begin with. The act of lifting the hand to reach the mouth with the bottle and tilting the neck backwards to drink tests hand to mouth coordination. This activity is at a higher difficulty level than 'Answering Phone' or 'Brushing hair' as it requires more precise placement and movement of the object in hand.

- **Ataxic:** Drinking requires precise positioning of the bottle and with hand and body shaking, subjects face extreme difficulty in drinking and end up in spilling.
- **Elbow Rigidity:** Subjects hold the bottle with a rigid, bent-elbow and rely only on shoulder movements to reach the mouth. As they struggle to reach the mouth with a rigid elbow they tend to move the neck forward to compensate for the lack of flexibility in the elbow.
- **Shoulder Weakness:** Similar to other activities the weaker shoulder droops and the head tilts towards the affected side while performing this activity.
- **Tremors:** While drinking water with tremors subjects find it difficult to hold the bottle steady at the mouth.

**Wearing glasses:** In this activity both the hands are involved. Subjects start in a position similar to 'Drinking' but hold the glasses with both the hands. In addition to test a subject's ability for hand to neck co-ordination, this activity requires better co-ordination between hands.

- **Ataxic:** As the whole body is shaking along with both hands, it takes considerable effort to put the glasses in place.
- **Elbow Rigidity:** To wear glasses with rigid, bent-elbows subjects rely on the shoulder to lift the hands up and towards the face. To compensate for the lack of flexibility in elbow subjects move their neck forward to wear the glasses.
- **Shoulder Weakness:** For this impairment subjects lift the unaffected arm and bend the affected arm from the elbow without lifting. This results in the glasses being held vertically and the subject has to tilt their head to wear them.
- **Tremors:** Like 'Drinking' subjects struggle to place the object (glasses) into the desired position.

**Brushing Floor:** This activity tests the functional independence and co-ordination of the whole body together. To perform the activity, subjects stand holding the brush and perform two forward strokes of the brush. While the arms perform the brushing strokes, the legs help in positioning. Thus, this activity requires functional independence and co-ordination of the whole body.

- **Ataxic:** With the whole body shaking along with both the arms subjects are barely able to perform the brushing strokes.
- **Elbow Rigidity:** Subjects start with bent, rigid elbows and use their shoulder to complete the brushing strokes. They also use their torso to move back and forth to compensate for the lack of flexibility in the elbows.
- **Shoulder Weakness:** For this impairment, subjects lift the brush with the unaffected arm. As with other activities subjects tend to tilt towards the weaker shoulder side.
- **Tremors:** Similar to 'Ataxic' subjects find it difficult to do proper strokes although in this case brushing strokes are better than 'Ataxic'.

## 5.4 Data Collection

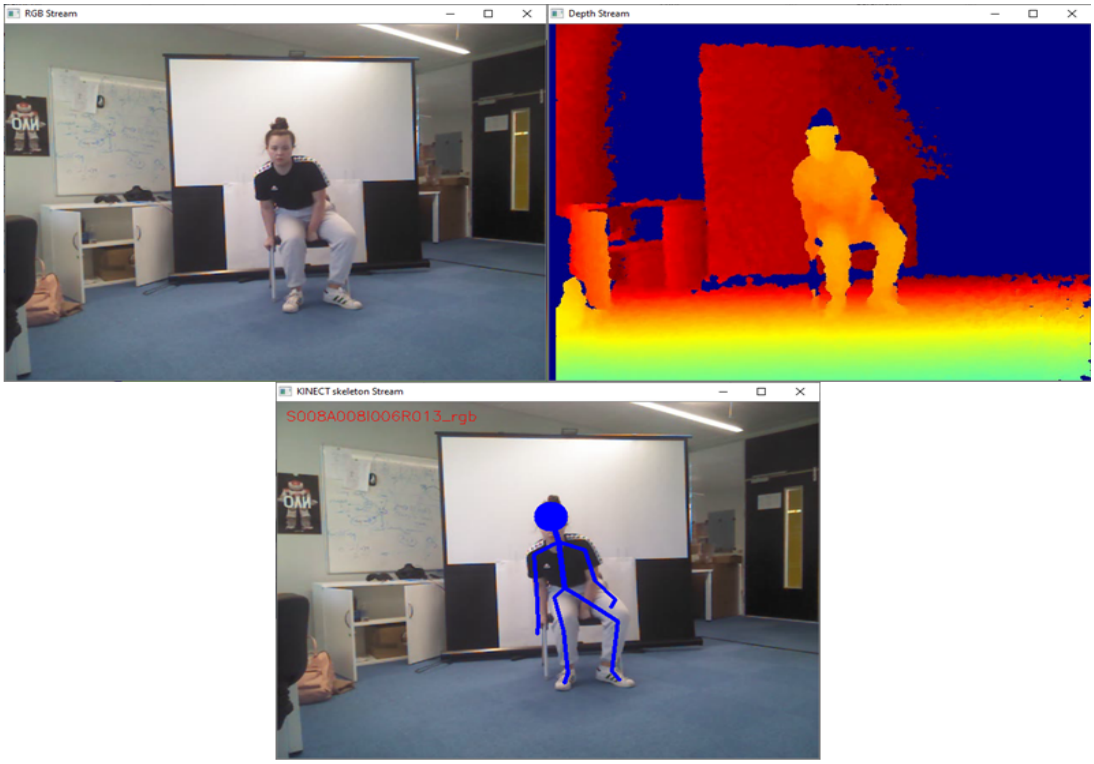



Figure 5.7: The dataset is captured through Kinect which captures the data in RGB, depth and 3D body-pose format. The raw depth data is storage-intensive and hence encoded in RGB format where different colours indicate different depths

The first step was to determine the dimensions of the dataset and as mentioned in Sec. 5.2, an initial goal of 5K samples was set. The intention was to present a dataset that contains equal distribution of samples across classes ('Activities' and 'Impairments') and subjects. This meant 100 samples for

each 'Activity-Impairment' combination of 10 ADL including 5 versions of each ADL. Many of the existing datasets (Table 3.1) have used 10 subjects and thus it was assumed that 10 subjects would provide a good enough inter-subject variation and generalisation. To test the feasibility of the plan in terms of time constraint, a pilot experiment was conducted. This included selecting a subject, introduction and ethical form signing, laboratory booking for filming and then the actual filming. While the actual filming took 10 hours, the whole process was completed over several sessions owing to availability of subject, availability of laboratory and required a month. The planned size of the dataset seemed suitable with regards to time constraints of the project.

Similar to majority of the datasets in CV-based assessment and rehabilitation (Table 2.7) as well as well-known datasets for human activity recognition (Table 3.1) Kinect (Su et al., 2014) sensor was used for filming. The Kinect device is capable of capturing data in RGB and Depth format. From the depth and the RGB data Kinect also calculates and provides estimates of human body pose in 3D format. However, Kinect does not provide a suitable software application that could be used to store RGB, Depth and Pose data. Therefore, a Python-based application was developed to store the filmed data. 'PyKinect 1.0' (Microsoft, 2012b) plugin was used as an interface to talk to the Kinect Driver called 'Kinect for Windows Runtime 2.0' (Microsoft, 2012a). RGB data was saved in 'avi' (Maertens; Soroushian, 2007) format while depth data was saved in raw format and pose data was saved as compressed Numpy arrays (Walt et al., 2011). The pose format requires PyKinect to translate the real-world coordinates into RGB coordinates.

Following design and application development, the next step was to select volunteers for filming. Drama students from Edge Hill University Performing Arts Department were invited and selected on a first come first serve basis. Drama students are well suited for such activities as they experience situations requiring them to enact various scenarios including acting as differently-abled. Ethical clearance was obtained from the Edge Hill University Ethical Committee (Appendix B.4). In line with ethical requirements, all the participants were given the Participant Information Sheet (PIS) and they signed the consent form (Appendix B.4). To protect identity, each participant was allocated a subject number. The number name mapping is safely stored in a locker and the subject number is being used for purposes like subject-wise train validation split and so on. It needs to be mentioned that subjects were informed (through PIS and verbally) that this dataset including the videos will be made publicly available upon the completion of this study and all the candidates agreed to it. Thus, while the data samples are not mapped to names or other details of the subject, the face will be displayed in the dataset. This is important as many pose-estimation approaches including Kinect depend on facial key-points such as nose for full body-pose estimation. Availability of face in the samples will help future users who may want to do pose estimation from the RGB and depth data through recent DL-based techniques (Pavllo et al., 2019; Chen; Ramanan, 2017). Eventually, a total of 10 subjects consisting of 6 female and 4 male subjects took part in the filming. Approximately 5.8K video sequences were filmed out of which 5685 were found to be correctly filmed sequence during post-processing.

| Activity | Code | Impairment | Code |
|---|---|---|---|
| Answering phone | A001 | Normal | I001 |
| Brushing floor | A002 | Ataxic | I002 |
| Brushing hair | A003 | Elbow rigidity | I003 |
| Clapping | A004 | Tremors | I004 |
| Drinking | A005 | Shoulder weakness | I005 |
| Reaching above | A006 | Knee rigidity | I006 |
| Sitting | A007 | Weakness to one side | I007 |
| Standing | A008 | Wider gait | I008 |
| Walking | A009 | | |
| Wearing glasses | A010 | | |

Table 5.1: 'Activity' and 'Impairment' codes

## 5.5 Post-processing and Statistics

h

First each RGB video was visualised with pose data superimposed on the RGB image, which helped to visually inspect the accuracy of the performed activity and the pose data. After careful consideration 5685 sequences were found to be usable out of around 5800 that were filmed. Then, the files and folder were renamed and arranged with the following structure:

$$\{Format\}\_data/Sss/Aaa/Iii/SssAaaIiiRrr\_\{Format\}.ext$$

Here $Format$ is either 'RGB', 'Depth' or 'Kinect'. 'RGB' indicates normal colour video while 'Kinect' indicates it is 3D body-pose data. 'Depth' indicates colour-coded depth information where raw depth data has been converted to RGB format with different colours indicating different depth. Thus, 'RGB' and 'Depth' files have the extension 'avi' whereas Kinect files have been saved as Numpy arrays in '.txt' format. 'Sss' indicate the subject ID, 'Aaa' the activity ID, 'Iii' the Impairment ID and 'Rrr' the repetition ID. The 'Activity' and 'Impairment' name to ID mapping for the dataset is given in Table 5.1. For example, the path and name of the $2_{nd}$ sequence (repetition) for answering phone with tremors from subject 5 would be:

$$rgb\_data/S005/A001/I007/S005A001I007R002\_rgb.avi$$
$$depth\_data/S005/A001/I007/S005A001I007R002\_depth.avi$$
$$kinect\_data/S005/A001/I007/S005A001I007R002\_kinect.txt$$

Now, the discussion presents some statistics to characterise the dataset further. Altogether the dataset presents 5685 samples in RGB, depth and Kinect-based 3D pose format which makes an average of 568.5 for the 10 activities. 'Reaching above' has the lowest number of sequences 559, while 'Drinking' has the highest number of sequences 579. With 100 sequences the Activity 'Standing' with Impairment 'Weakness to One Side' has the lowest number of sequences. The highest number of sequences is 118 for 'Clapping-Tremors', 'Drinking-Tremors', 'Drinking-Shoulder Weakness' and

'Brushing Floor-Knee Rigidity'. The average for each 'Activity-Impairment' combination is 113.7, which is comfortably more than 100 targeted initially. The 'Normal' version and the 'Ataxic' impairment is available for all the ADL and thus have more samples than the other impairments. The comparison between number of sequences filmed for each 'Activity-Impairment' combination is illustrated graphically in Figure 5.8. Please refer to Appendix A.1 for further details on the number of sequences filmed for each activity, impairment and subject.

| Impairments-> Activities | Normal | Ataxic | Weakness to One Side | Wider Gait | Knee Rigidity | Elbow Rigidity | Tremors | Shoulder Weakness | Total |
|---|---|---|---|---|---|---|---|---|---|
| Walking two steps | 114 | 118 | 114 | 113 | 113 | NA | NA | NA | 572 |
| Sitting | 111 | 116 | 111 | 116 | 113 | NA | NA | NA | 567 |
| Standing | 107 | 115 | 100 | 113 | 112 | NA | NA | NA | 547 |
| Clapping | 115 | 116 | NA | NA | NA | 112 | 118 | 116 | 577 |
| Reaching Above | 113 | 106 | NA | NA | NA | 119 | 117 | 104 | 559 |
| Brushing Hair | 111 | 117 | NA | NA | NA | 116 | 112 | 115 | 571 |
| Answering Phone | 109 | 117 | NA | NA | NA | 115 | 111 | 114 | 566 |
| Drinking | 111 | 116 | NA | NA | NA | 116 | 118 | 118 | 579 |
| Wearing Glasses | 106 | 116 | NA | NA | NA | 116 | 113 | 120 | 571 |
| Brushing Floor | 115 | 120 | NA | NA | NA | 118 | 117 | 106 | 576 |
| Total | 1112 | 1157 | 325 | 342 | 338 | 812 | 806 | 793 | 5685 |

Table 5.2: Dataset details highlighting the number of sequences obtained for each 'Activity' and 'Impairment' combination

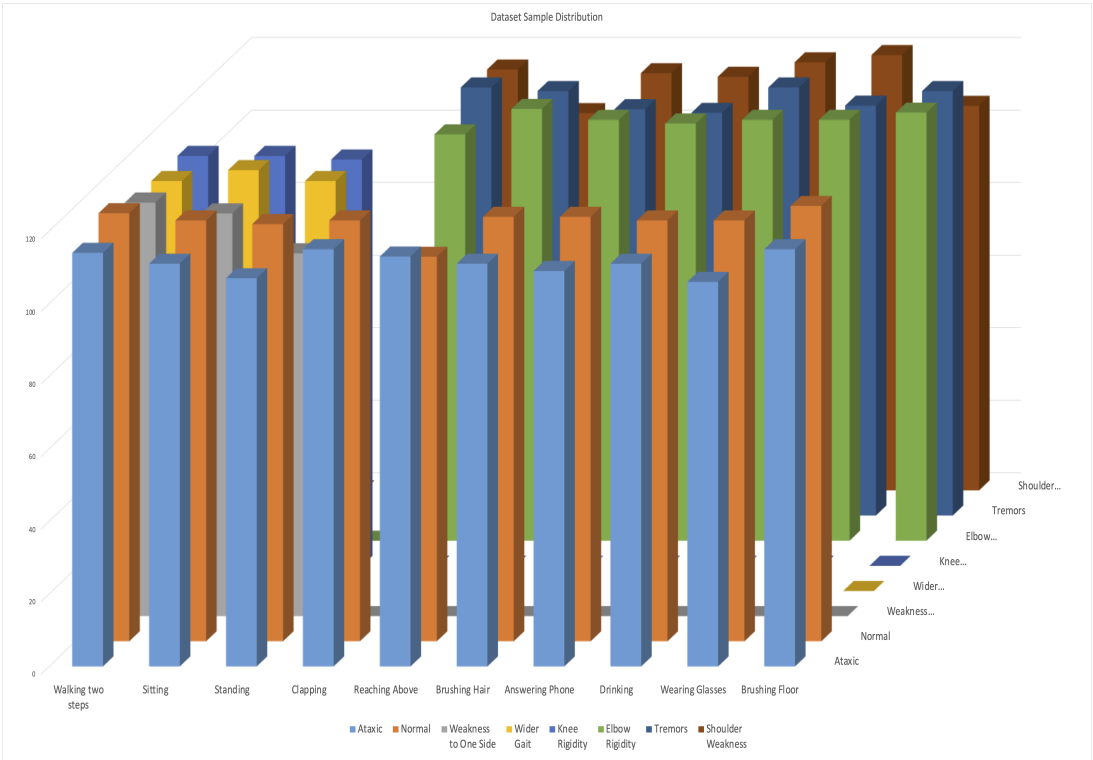



Figure 5.8: Dataset details highlighting number of sequences obtained for each 'Activity' and 'Impairment' combination graphically

Figure 5.9 illustrates the subject-wise distribution of sequences present in the dataset. The initial goal was to film 500 sequences for each subject for a total of 5000 sequences. As discussed in the previous section, to account for filming error more sequences were filmed which totalled around 5800 videos. Thus, the initial goal of having 500 sequences per subject was comfortably achieved. The variation in distribution is due to variation in correct number of sequences that could be obtained from each subject.

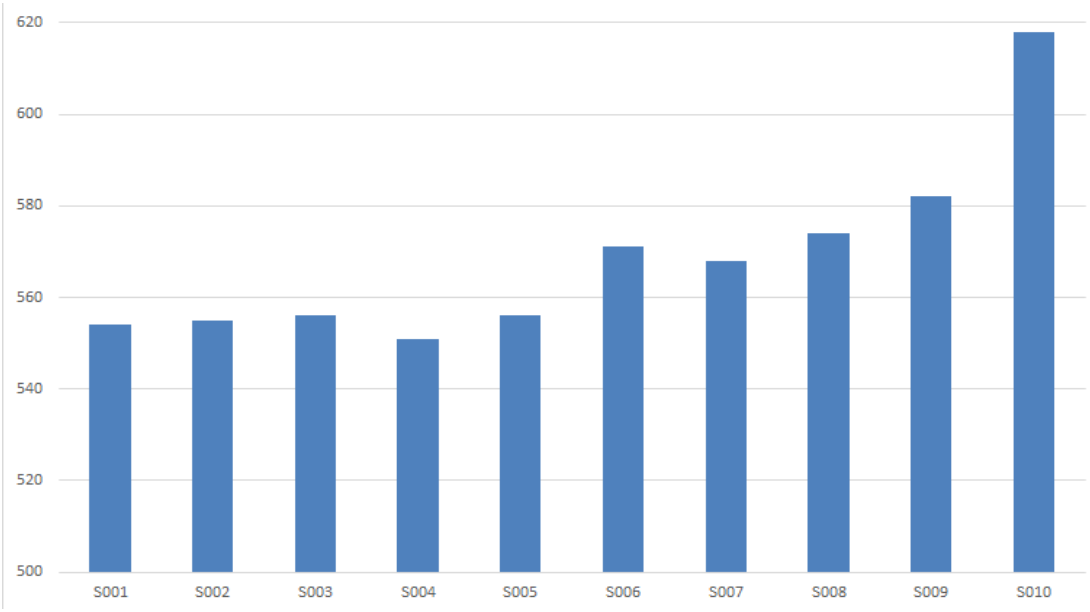



Figure 5.9: Subject-wise distribution of samples. X-axis: Subject ID, Y-axis: Number of samples

The next figure (Figure 5.10), shows the distribution of the number of frames present in each of the 5685 sequences. The high variation in distribution points to varying nature of 'Activities' and 'Impairments' involved. For example, the 'Answering Phone' with 'Tremors' is a much slower action than 'Standing' with 'Normal'

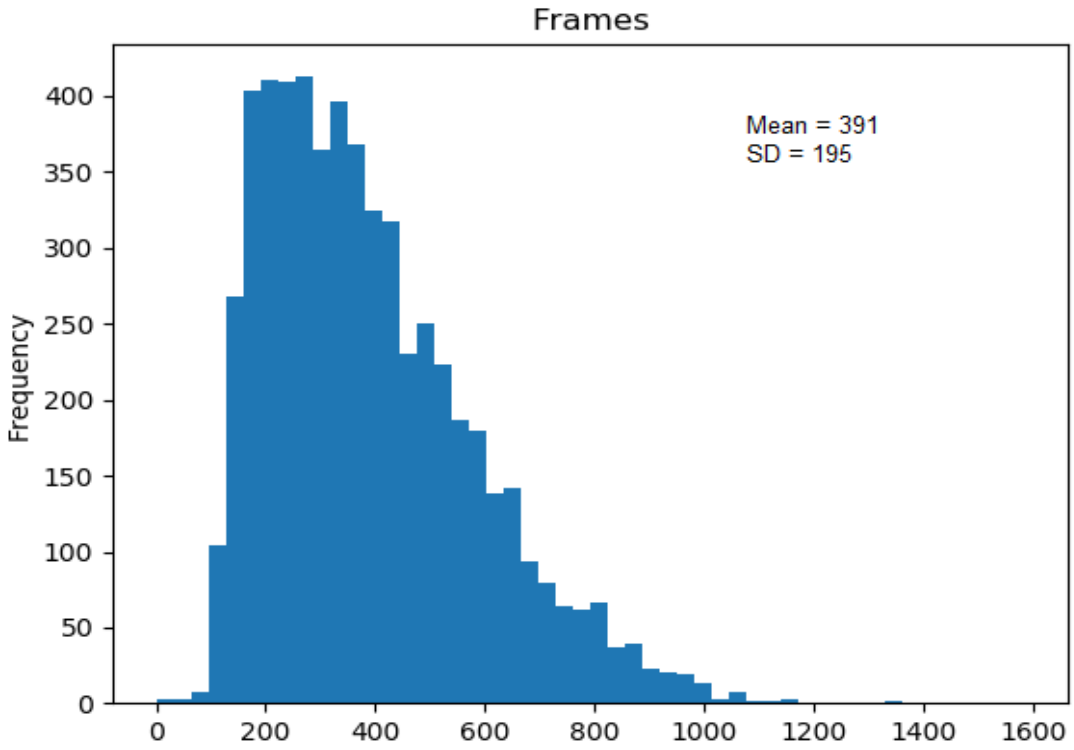



Figure 5.10: Frame distribution

## 5.6 Discussion

The dataset presented in this chapter differentiates between ADL performed by a healthy individual to that of a physically impaired person. In this regard it is not only important to illustrate the difference, but it is quintessential to accurately represent the physically impaired version of the same ADL. The pattern and extent of deviation from normal ADL depends upon the physical impairment. An inaccurate representation of the physical impairment specific ADL may lead to miss-classification of the impairment. For reasons mentioned in Sec. 5.2, it was not feasible to involve actual patients for the data collection. To ensure that the filmed samples accurately represent ADL as performed by actual patients several measures were taken. First, as mentioned in Sec. 5.2, an Occupational Therapist provided video demonstration of ADL as performed by physically impaired persons. Then, Drama students were specifically selected to act as physically impaired persons. Drama students perform acting as part of their day-to-day learning and are well suited to adapt to such situations. Finally, a few samples were randomly reviewed with the Occupational Therapist to ensure that filmed samples accurately represented the video sequences used as guidance. However, it needs to be noted that due to nature of the human body, the extent of impairment and other physiological factors, the same impairment can be exhibited very differently in different persons. The dataset only captures a limited set of scenarios for each ADL and its corresponding impairments. For example, while 'Clapping hands' with 'Rigid Elbow', both elbows are locked in their position which is not often the case. Similarly, 'Drinking Water' with 'Elbow Rigidity' is performed with the right hand only which may not be the case always. In the dataset, 'Elbow Rigidity' and 'Knee Rigidity' has been illustrated as fully locked knee or elbow respectively. In reality, the rigidity differs from case to case. Any dataset meant for practical or commercial life application would need to capture all these scenarios. People

normally perform ADL to achieve some functional objective like 'Drinking', 'Brushing' and others. Thus, the ADL chosen for this dataset are also functional in nature. However, in normal day-to-day life people perform these activities as a part more complex activity. For example, 'Reaching Above' in kitchen can be performed as a part of cooking. In such day-to-day scenarios 'Reaching Above' maybe performed in multiple different ways. This also may be in combination with other activities. For example, while cooking a person may try to 'Reach Above' for grabbing something, putting something or clean something. The person may reach above and carry out some tasks for an extended period of time. Clearly, this is very different from the rather simplistic approach of the dataset where the person just 'Reaches Above'. The dataset is not targeted towards such realistic day-to-day scenarios. Rather, the objective of this dataset targeted towards point in time functional assessment of physically impaired persons. In such situations Health Carers may ask patients to perform simplified ADL tasks as illustrated in the dataset, for assessment.

This section also presents a comparison with existing datasets and discusses the potential impact on future research in this area. The dataset is potentially applicable to two areas of research. First is in the domain of CV-based rehabilitation and assessment. The review presented in Chapter 2 (Sec. 2.9) shows that there are very few publicly available datasets that present physically impaired persons' activities. Table 2.7 shows that currently existing datasets in this domain are very small for evaluating today's DL models. These datasets mostly consist of movements limited to one or few specific activities (e.g., 'Sit to Stand') or body parts. Like other areas in CV, CV-based rehabilitation and assessment methods have seen increasing use of DL-based methods. But, due to lack of large generalised publicly available datasets in this domain, it is yet to be fully explored by the CV and AI community. The proposed dataset aims to encourage DL or AI researchers to explore and contribute towards this domain. The dataset is much larger than other datasets compared in Table 5.3. Also, the proposed dataset contains whole body movements while the other datasets address only specific body parts and lack generalisation. To the best of my knowledge, this is the first dataset that presents both normal and physical-impairment specific versions of ADL and thus presents a larger, more generalised approach for evaluating the functional ability of a patient.

| Author | Impairment | #Videos | #Subjects | Remarks |
|---|---|---|---|---|
| SPHERE-Staircase2014 (Paiement et al., 2014) | Walking-up stairs | 48 | 12 subjects | Normal and abnormal gait |
| SPHERE-Walking2015 (Tao et al., 2016) | Walking | 40 | 10 | normal and abnormal gait |
| SPHERE-SitStand2015 (Tao et al., 2016) | Sit to stand | 109 | 10 | Restricted knee, hip, freezing |
| TRSP (Dolatabadi et al., 2017) | Stroke | NA | 20 | 4 compensatory movements |
| Parkinson's pose estimation (Li et al., 2018b) | PD, LID | 526 | NA | 4 UPDRS assessment tasks |
| UI-PRMD (Vakanski et al., 2018) | General exercises | 100 | 10 | Rehab Exercises |
| KIMORE Dataset (Capecci et al., 2019) | Stroke, PD | 1950 | 78 | 5 exercises |
| AHA-3D Dataset (Antunes et al., 2018) | Lower body abilities | NA | 21 | 4 exercises |
| Proposed Dataset | ADL | 5685 | 10 | 4 impairments for each ADL |

Table 5.3: The proposed dataset in comparison to publicly available datasets aimed towards CV-based rehabilitation and assessment

The second area of research to which this dataset can potentially contribute is CV and AI-based human activity recognition. In the domain of object detection and recognition, many datasets present multi-label targets. For example, Russakovsky et al. (2015) present multiple object attributes such as colour, shape, pattern and texture. But multi-label activity recognition is yet to fully explored by the CV community. Table 5.4 shows that none of the currently existing datasets have explored the area of multi-label activity recognition. Multi-label object detection has generated great attention in the AI and CV community and has vastly helped progression in this area with concepts like Mask RCNN (He et al., 2017). Similarly, one can hope that the proposed dataset will greatly advance the research in multi-label activity recognition, which is yet to be explored by the AI and CV community.

| Datasets | #Videos | #Classes | #Sub-classes | #Subjects | Data Modalities |
|---|---|---|---|---|---|
| MSRDailyActivity3D (Wang et al., 2012) | 320 | 16 | 0 | 10 | R,D,3J |
| UTKinect (Xia et al., 2012a) | 200 | 10 | 0 | 10 | R,D,3J |
| MSR-Action3D (Li et al., 2010) | 567 | 20 | 0 | 10 | R,D,3J |
| CAD-60 (Sung et al., 2011) | 60 | 12 | 0 | 4 | R,D,3J |
| CAD-120 (Koppula et al., 2013) | 120 | 20 | 0 | 4 | R,D,3J |
| NTU-RGBD (Shahroudy et al., 2016) | 58K | 60 | 0 | 40 | R,D,3J |
| Northwestern-UCLA (Wang et al., 2014a) | 1475 | 10 | 0 | 10 | R,D,3J |
| Proposed Dataset | 5865 | 10 | **10** | 10 | R,D,3J |

Table 5.4: Comparison of the proposed dataset with other activity recognition datasets. R: RGB, D: Depth, J: Joint

## 5.7 Conclusion

This Chapter addresses the third objective of the current study, which is to prepare a dataset that captures normal as well as physical-impairment specific versions of ADL. The dataset presents 5685 sequences of 10 subjects, performing 10 different ADL with 7 different impairments in total. For each ADL, there is one healthy four physical impairment-specific versions performed by healthy subjects acting like patients guided by an Occupational Therapist. The dataset has the potential to further the research on CV-based rehabilitation and assessment. To the best of my knowledge,

this is the first dataset that presents ADL and include the respective physical impairment-specific versions. In the area of CV and AI-based human activity recognition this dataset can pave the way for multi-label activity recognition. The plan is to release this dataset publicly upon the completion of this study. In the next two chapters, DL-based human activity recognition models are presented which contribute towards the multi-label activity recognition method presented in Chapter 8. The multi-label activity recognition model presented in Chapter 8 has been trained and evaluated using the current dataset. The dataset along with the model presented in Chapter 8 has been submitted to the *IEEE International Conference on IROS, 2021.*

# Chapter 6

# Human Activity Recognition: Model 1

## 6.1 Introduction

This Chapter caters to the fourth objective of this study, which is to contribute a novel DL-based human activity recognition model. The main aim of the research is to contribute a novel model that can not only recognise an ADL, but also discriminate the impairment-specific variations of the same ADL. To this end, a multi-label activity recognition dataset was presented in the previous Chapter that contains normal ADL as performed by healthy subjects as well as four different impairment-specific versions of the same ADL. The dataset contains two labels for each sample ('Activity' and 'Impairment') and the next task is to prepare a model that can recognise these 'Activities' as well as 'Impairments'. The study approaches this task by first focusing on human activity recognition where the goal is to only recognise different ADL. This means there is only one-label ('Activity') for each sample. The area of ADL recognition has been extensively explored by the CV and AI community (Vrigkas et al., 2015) and relevant advances in this field has been duly described in the literature review (Chapter 3, Sec. 3.4). From the literature review it is clear that for comparative evaluation of any model one needs to use well-known benchmark datasets. Thus, in this Chapter and the next, two novel DL-based human activity recognition methods are presented that are evaluated on well-known publicly available datasets. This Chapter presents an *'Attention-based Learn-able Pooling'* method for human activity recognition from monocular RGB video data. First, the model uses an ImageNet (Jia Deng et al., 2009) dataset pre-trained Inception-ResNet-V2 (Szegedy et al., 2017) CNN network, which is well-known for its spatial processing capabilities. Then, a 'Self-Attention' mechanism followed by a Bi-LSTM is used to improve the network's temporal processing capabilities. This is followed by an activity-aware learn-able pooling mechanism based on FV, that exploits the temporal structures and dependencies contained in the Bi-LSTM's hidden states. To the best of my knowledge, this is the first approach that attempts to exploit temporal structures contained within a Bi-LSTM's hidden states by integrating semantic clustering (with FV) within a deep network.

The next section describes the rationale behind the proposed approach which is followed by two sections that revisit the Inception-ResNet-V2 (Szegedy et al., 2017) and the FV (Perronnin; Dance, 2007) that forms the basis of this model. The subsequent sections discuss the proposed approach and the experiments along with analysis of results. This is followed by a discussion on the proposed model highlighting its impact on the current research and on the broader research area.

### 6.1.1 Motivation/Rationale

Recent activity recognition models have focused on multiple modalities like CV-based 3D human body poses, RGB videos and depth maps. Due to the popularity of depth-based 3D pose estimation devices (e.g., Microsoft Kinect), and relatively less memory requirements of pose data, authors have increasingly relied on pose-based methods (Kim; Reiter, 2017; Xu et al., 2018). However, depth-based body pose estimation devices often suffer from inherent inaccuracies (Galna et al., 2014) and require both RGB and depth information resulting in processing of a high volume of data, which is computationally expensive. Moreover, despite the popularity of RGB-D devices, monocular RGB cameras are widely used in situations such as CCTV surveillance, home monitoring, etc. Therefore, in this Chapter a model purely based on monocular RGB videos is presented. An RGB video contains a lot of information regarding the scene, as well as objects handled by subjects, and can provide contextual information, which is vital in discriminating various human activities. As described in the literature review (Chapter 2, Sec. 3.4), this has been explored by deep CNNs, resulting in higher recognition accuracy (Baradel et al., 2018b; Sharma et al., 2016). The review also shows that existing approaches often combine the spatial information (from a CNN) (Ma et al., 2016) and temporal dependencies by using recurrent networks such as LSTMs (Baradel et al., 2018a). Thus, the current model first process spatial information using Inception-ResNet-V2 (Szegedy et al., 2017) network, which is well-known for its spatial processing capabilities followed by a Bi-LSTM-based temporal processing network.

Modern deep CNNs capture hierarchical feature representation of a given image. Its prediction is dominated by the task-specific representation of convolutional layers. These models have shown remarkable success in visual recognition by considering full images with distinctive classes. However, it raises questions about their performance in discriminating small changes in successive frames in a given video. Therefore, there is a need for learning meaningful spatio-temporal structures in videos for discriminating various human activities. To address this, a novel learn-able pooling mechanism is used, which captures the activity-aware spatio-temporal structure in videos by exploring both spatial and temporal information. The spatial information is explored using the high-level frame-wise features from the pre-trained CNN model Inception-ResNet-V2 (Szegedy et al., 2017). The dynamics of these spatial features over a given sequence and their importance for a given activity is captured using a Bi-LSTM and a novel 'Attention' mechanism, which captures both sequential and spatial 'Attention' by focusing on various temporal and spatial locations in the sequence. The novel 'Attention' mechanism presented in this Chapter consists of two parts: 1) A novel 'Sequential Self-Attention' mechanism to selectively focus on important temporal points by using high-level frame-wise CNN features; 2) The output of this 'Sequential Self-Attention' is fed into a Bi-LSTM to capture the long-term temporal dependencies, which is captured by its hidden states. To guide the model to discriminate the subtle changes in videos, a learn-able pooling method is proposed to capture the structural information and similarities contained within the Bi-LSTM's hidden states. To achieve this, a FV representation, which is based on a clustering mechanism has been adapted to semantically group information in hidden Bi-LSTM states. By exploiting the Bi-LSTM's hidden states with learn-able FVs, the model is able to select the hidden states based on their discriminative abilities. The output of learn-able FVs is pooled using activity-aware pooling to represent the

number of states equalling the number of activity classes. The novel learn-able FVs with activity-aware pooling replaces the customary GAP and FC layers towards the end. As a result, the proposed model is very flexible and can be adapted to any existing CNN model. The next two sections revisit the Inception-ResNet-V2 CNN model and FV which forms the basis of the current model.

## 6.2 Inception-ResNet-V2

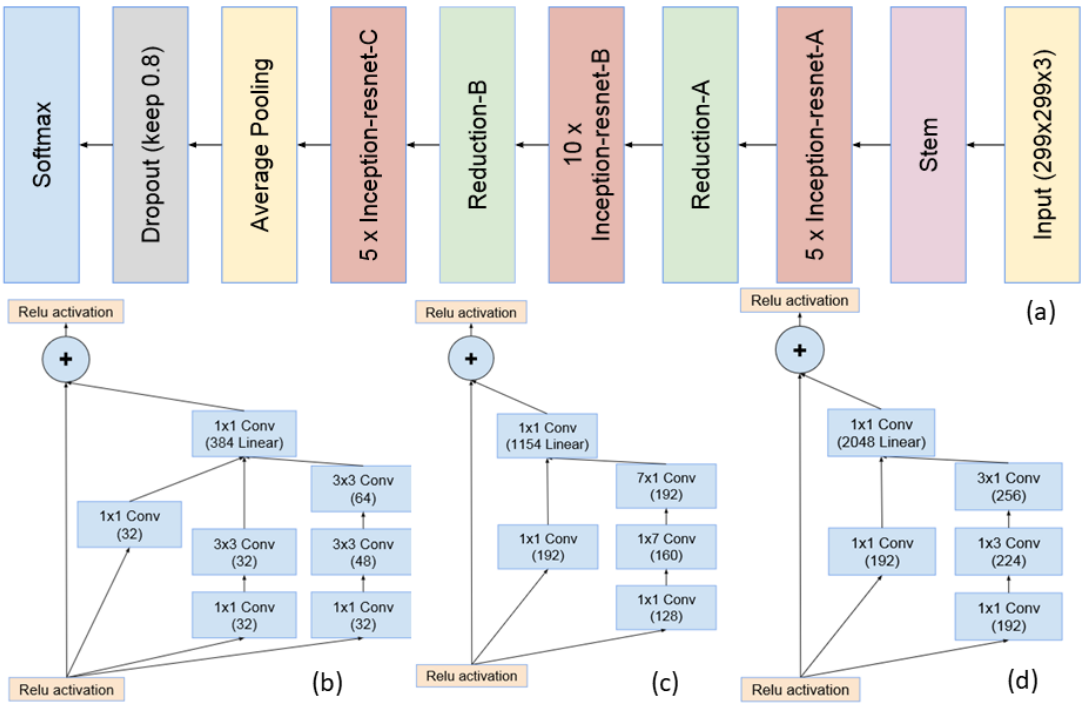



Figure 6.1: The current model is based on Inception-ResNet-V2. a) Overall architecture b) Inception-ResNet-A c) Inception-ResNet-B d) Inception-ResNet-C. The Figure has been referenced from Szegedy et al. (2017)

Inspired by the success of Inception blocks (Szegedy et al., 2015) and the impact of residual connections (He et al., 2016), Szegedy et al. (2017) proposed the Inception-ResNet-V2 architecture. In this architecture, authors combine Inception blocks with residual connections as shown in Figure 6.1. The core of the network is composed of three sequentially placed Inception-Residual blocks followed by spatial reduction (Figure 6.1a). Combined with increasing number of filters from top to bottom, spatial reduction is used by many standard CNN architectures (Howard et al., 2017; Szegedy et al., 2015; Szegedy et al., 2016). The output of the final Inception block is 1536 maps of spatial extent 8x8. These maps are passed through a GAP layer which pools the spatial dimensions from 8x8 to 1. A final dense layer with Soft-max activation function is applied for classification. The main novelty of this network lies in the Inception-ResNet blocks (Figure 6.1b, c, d). According to the authors, the goal of the Inception blocks is to reduce the number of parameters while maintaining efficiency and performance. Reduced number of connections means, faster and more efficient network with less over-fitting. Thus, instead of connecting all the filters together as previously done by VGG-16 (Simonyan; Zisserman, 2014), AlexNet (Krizhevsky et al., 2012), connections are modularised in the form of Inception blocks. In a layer (e.g., Inception-ResNet-C, Figure 6.1a) there are number of Inception blocks (in this case 5) which are not mutually connected but the outputs are concatenated at the end of the layer. Further, reduction in parameters is made through the use of $1 \times 1$ convolution

and factorised convolutions. In Inception-ResNet-A (Figure 6.1b), instead of using one 5 convolutions two $3 \times 3$ convolution layers are used. A layer of 5 convolutions has 25 parameters, whereas two layers of $3 \times 3$ convolutions have 18 ($2 \times 3 \times 3$) parameters and thus gives a reduction of 28%. The other form of factorised convolution is called asymmetric convolution. This type of convolution is used is the Inception-ResNet-B (Figure 6.1b) and the Inception-ResNet-C (Figure 6.1c) block. In these blocks, instead of using a $3 \times 3$ or $7 \times 7$ convolution $3 \times 1$, $1 \times 3$ and $7 \times 1$, $1 \times 7$ convolution is used respectively. A $7 \times 7$ convolution has 49 parameters whereas asymmetric factorisation leads to $7 \times 1 + 1 \times 7 = 14$ parameters only. Owing to the amount of reduced connections, the network becomes more robust to over-fitting and can go deeper to improve the performance.

## 6.3 Fisher Vector

The FV implemented here is adapted from Sánchez et al. (2013) and everything described below is from the same article. In statistics, the definition of a score function is the gradient of the log-likelihood of the data on the model:

$$G_{\lambda}^{X} = {}_{\lambda} \log u_{\lambda}(X) \tag{6.1}$$

Here, $X = \{x_t, t = 1, ..., T\}$ is a D-dimensional local descriptor extracted from image or video features. Fisher kernel $K_{FK}$ measures the similarity between two such data points say X and Y:

$$K_{FK} = G_{\lambda}^{X'} F_{\lambda}^{-1} G_{\lambda}^{Y} \tag{6.2}$$

Here, $F$ is the Fisher Information Matrix (FIM). $F$ is positive semi-definite, thus can be Cholesky decomposed as $F^{-1} = L_{\lambda}^{'} L_{\lambda}$. FV is defined as the sum of normalised gradient statistics and is formally defined as:

$$FV_{\lambda}^{X} = \sum_{t=1}^{T} L_{\lambda\lambda} \log u_{\lambda}(x_t) \tag{6.3}$$

To put it simply, FVs adapted to image classification are nothing but gradients of parameters from GMM (Titterington et al., 1985). Thus, in Eq. 6.1, $\lambda$ corresponds to the parameters of GMM, which are the mixture weight $w_k$, the mean vector $c_k$ also called cluster centre and the co-variance matrix $\Sigma_k$ of the Gaussian $k$. Here, $k$ is the number of components or clusters in the GMM. $u_{\lambda}$ in Eq. 6.1, when defined in terms of GMM is given by:

$$u_{\lambda}(x) = \sum_{k=1}^{K} w_k u_k(x) \tag{6.4}$$

The $k_{th}$ Gaussian for data point $x$ is defined as $u_k$:

$$u_k(x) = \frac{1}{(2\pi)^{D/2} \mid \Sigma_k \mid^{1/2}} \exp\{-1/2(x - c_k)' \Sigma_k^{-1} (x - c_k)\} \tag{6.5}$$

and it is required that all the mixture weights add up to 1:

$$\forall : w_k \geq 0, \sum_{k=1}^{K} w_k = 1 \tag{6.6}$$

The weights are adapted in a soft-max manner as done by Krapac et al. (2011), which is defined as:

$$w_k = \frac{\exp(\alpha_k)}{\sum_{j=1}^{K} \exp(\alpha_j)} \tag{6.7}$$

Further, the weights can be combined with $k_{th}$ Gaussian for each data point $t$ to give the soft assignment of point $x_t$ to the Gaussian $k$. This is defined by $\gamma_t(k)$, which is also called the posterior probability:

$$\gamma_t(k) = \frac{w_k u_k(x_t)}{\sum_{j=1}^{K} w_j u_j(x_t)} \tag{6.8}$$

The authors assume $\Sigma_k$ as a diagonal co-variance matrix which is denoted by $\sigma^2$. The soft-max formalism avoids explicitly enforcing the constraint of Eq. 6.6. Thus, the final parameters of GMM are $\lambda = \{\gamma_t, c_k, \sigma_k, k = 1...K, t = 1...T\}$. As a result, the gradients of a single descriptor $x_t$, from Eq. 6.1 w.r.t the parameters $\lambda$ are:

$$\Delta_{ck} \log u_\lambda(x_t) = \gamma_t(k) - w_k \tag{6.9}$$

$$\Delta_{\alpha_k} \log u_\lambda(x_t) = \gamma_t(k)(\frac{x_t - c_k}{\sigma_k^2}) \tag{6.10}$$

$$\Delta_{\sigma k} \log u_\lambda(x_t) = \gamma_t(k)(\frac{(x_t - c_k)^2}{\sigma_k^3} - \frac{1}{\sigma_k}) \tag{6.11}$$

To obtain FVs as described in Eq. 6.3, the above equations need to be combined with $L_\lambda$. To compute $L_\lambda$, which is the inverse of the square root of FIM, the authors make the following assumption. The soft-assignment distribution $\gamma_t(i)$ (Eq. 6.8) sharply peaks on a single value of any descriptor $x_t$. Under this assumption, FIM is diagonal which gives the following equations:

$$FV_{\alpha_k}^X = \frac{1}{w_k^{1/2}} \sum_{t=1}^{T} (\gamma_t(k) - w_k) \tag{6.12}$$

$$FV_{c_k}^X = \frac{1}{w_k^{1/2}} \sum_{t=1}^{T} \gamma_t(k)(\frac{x_t - c_k}{\sigma_k}) \tag{6.13}$$

$$FV_{\sigma_k}^X = \frac{1}{w_k^{1/2}} \sum_{t=1}^{T} \gamma_t(k)[\frac{(x_t - c_k)^2}{\sigma_k^2} - 1] \tag{6.14}$$

The above equations are derived from the approximation that FIM is diagonal and is left out for simplicity. In the current study, Eq. 6.12 to Eq. 6.14 are adapted and integrated to the DL-based model for learn-able pooling with FVs.

## 6.4 Proposed Approach: ADL Recognition Model 1

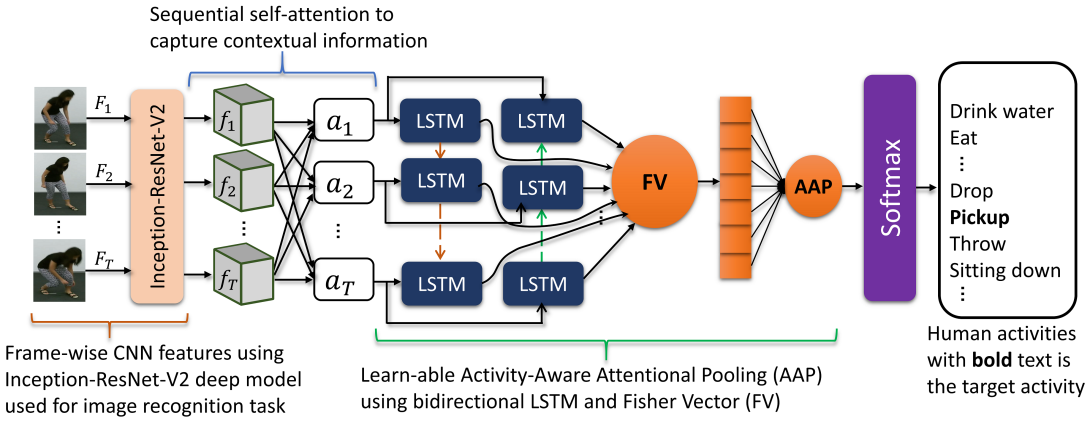



Figure 6.2: The proposed deep network consists of: 1) A pre-trained CNN (Inception-ResNet-V2 (Szegedy et al., 2017)) model used to extract frame-wise high-level CNN features from a given video consisting of $T$ frames. 2) A 'Sequential Self-Attention' layer to capture the contextual information consisting of important spatial and temporal knowledge. 3) Learn-able activity-aware pooling consisting a Bi-LSTM and FV to learn the structural information and similarities by exploring the hidden states of the Bi-LSTM. The Bi-LSTM is unrolled to illustrate its hidden states for the video of duration $T$. The activity aware feature vector is passed through the Soft-max layer to estimate the probabilities of various human activities

The major components of the proposed network are shown in Figure 6.2. It uses the Inception ResNet-V2 (Szegedy et al., 2017) to extract the frame-wise CNN features. Inception ResNet-V2 is well-known for its impressive performance in solving image classification and object detection problems. It is used in a time distributed manner in which all the frames from a given video are passed through the same layers to extract the corresponding CNN features. These features are then processed by the adapted 'Sequential Self-Attention' mechanism to capture the contextual information consisting of important temporal knowledge. It captures the information describing how much to recommend the CNN features at time point $t$ in focus, conditioned on all other CNN features from different time points. Afterwards, the Bi-LSTM FV-based learn-able pooling method is used which enhances the network's ability to comprehend long-term temporal structures and dependencies. This is done by exploiting the structural information contained in the Bi-LSTM cells by semantically grouping its hidden states into learn-able clusters, which are part of the FV representations. The literature review (Chapter 3, Sec. 3.2.3) shows researchers' recent inclination towards learn-able pooling approaches (Girdhar et al., 2017; Miech et al., 2017; Arandjelovic et al., 2016) as compared to statistical pooling (e.g., GAP, max-pooling, etc.) to pool the most relevant features. Typically, such methods have two parts: i) Calculating a learned representation and ii) A weighted pooling method. The learn-able pooling method used in the model is inspired by the NetFV (Miech et al., 2017) which uses i) FV and ii) first-order weighted pooling. The current study's adaptation of FV is different from NetFV in the following ways:

- NetFV does not take into account the temporal information contained in the video frames whereas the proposed model learns FV from temporal information contained in a Bi-LSTM's hidden states.
- NetFV uses a FC layer towards the end whereas the current study uses first-order activity-aware

pooling as the final classification layer.

In NetFV, authors learn FV directly from CNN features for processing video information and therefore, do not consider any temporal information. This study adapts 'Attention' weighted CNN features processed through a Bi-LSTM. This helps to exploit the temporal structure contained within the Bi-LSTM's hidden states and grouping them in a semantic manner. In NetFV (Miech et al., 2017), the pooled size is a tune-able hyper-parameter which necessitates further layers for classification. But in activity-aware pooling mechanism the pooling weights itself act as the final classifier. In other words, the number of semantic clusters produced by FV is equal to the number of activity classes.

### 6.4.1 Problem Formulation

In the video-based human activity recognition, a set of videos $V = \{v_1, v_2, \ldots, v_N\}$ and their respective activity labels $Y = \{y_1, y_2, \ldots, y_N\}$ are provided, where $N$ is the total number of videos. The objective is to find a mapping function $\mathcal{F}$ that predicts $\hat{y} = \mathcal{F}(v)$ which matches the actual activity $y$ of a given video $v$ as much as possible. The ultimate aim is to learn $\mathcal{F}$ by minimising the categorical cross-entropy $E_v$ between the predicted activity label $\hat{y_n}$ and the actual label $y_n$ via a training procedure. The entropy $E_v$ is computed as:

$$E_v = -\sum_{n=1}^{N} y_n \log(\hat{y}_n), \text{ where } \hat{y}_n = \mathcal{F}(v_n) \tag{6.15}$$

RGB videos contain both spatial and temporal information that needs to be accurately represented to learn $\mathcal{F}$. It is well-known that CNNs are widely used for capturing the spatial information in solving visual recognition tasks. Let's say a video $v = \{F_1, F_2, \ldots, F_T\}$ consists of $T$ frames and the first objective is to extract visual features $f = \{f_1, f_2, \ldots, f_T\}$ for the respective frames in video $v$ as shown in Figure 6.2. This is achieved by using any existing CNN model used for image or object recognition task.

### 6.4.2 Spatial Features Extraction

In this model, Inception-ResNet-V2 (Szegedy et al., 2017) is used in a pre-trained manner (trained on ImageNet (Russakovsky et al., 2015)) to extract the frame-wise high-level CNN features. For a given video frame $F_t$ of size $299 \times 299 \times 3$ (width = height = 299 and 3 channels representing RGB), the corresponding output $f_t$ of the final Inception block of the network is $8 \times 8 \times 1536$ feature map with a spatial extent of $8 \times 8$. These maps are then passed through a GAP layer, which pools the spatial dimensions from $8 \times 8$ to 1 and provides a final feature vector of size 1536. This is used in the next step of 'Sequential Self-Attention' in capturing contextual information as shown in Figure 6.2. The network is applied in a time distributed manner for processing video inputs. In time distributed processing, the frames are individually processed but the network weights are shared across frames. The output of Inception-ResNet-V2 is fed to a 'Sequential Self-Attention' method which is discussed next.

### 6.4.3 Temporal Processing to Capture Contextual Information

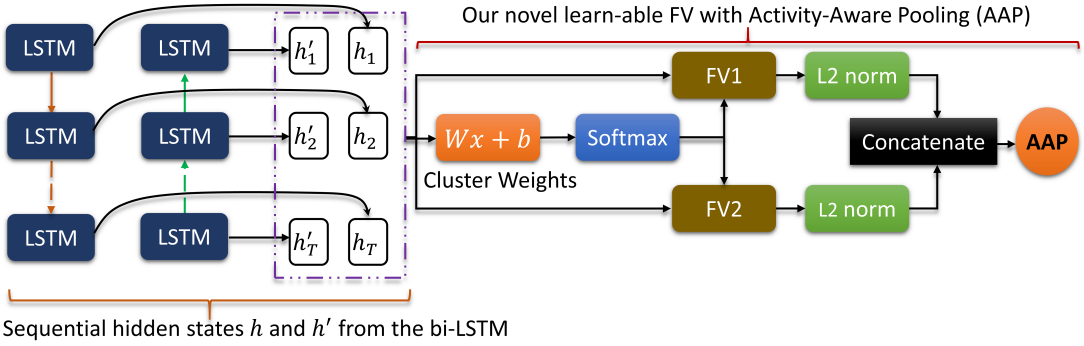



Figure 6.3: The proposed learn-able FV pooling using a Bi-LSTM: The structural information in hidden states of the Bi-LSTM is learned through FV. For clarity, the Bi-LSTM is unrolled to illustrate the hidden states over the video duration of $T$. The FV cluster weights are learned through weight matrix $W$ and bias $b$. The weights are then used for deriving the first order ($FV_1$) and the second-order ($FV_2$) FV. The FVs ($FV_1$ and $FV_2$) have learned parameters consisting of cluster centres and co-variances as shown in Eq. 6.19. Towards the end, $FV_1$ and $FV_2$ is concatenated and pooled with activity-aware weighted pooling for human activity classification

To capture the contextual information from the sequence of feature map $f_t$ $(t = 1 \dots T)$ as outputs from the Inception ResNet-V2, 'Sequential Self-Attention' mechanism is used that transforms the feature map into a weighted version of itself, conditioned on the rest of the feature maps representing the remaining frames. This leads the network to selectively focus on more relevant features to generate holistic context information for further processing by learn-able pooling for activity recognition. The 'Attention' mechanism helps to focus the network on more relevant features for discrimination while LSTM helps to represent the long-term temporal dependencies. More specifically this study explores the use of 'Self-Attention' (Zhang et al., 2018) for activity recognition. The goal of the 'Attention' mechanism is to assign a higher weight to more relevant features. Normally, 'Attention' mechanism is described as a mapping function. It maps a query $\mathcal{Q}$ and a set of key-value pairs $\mathcal{K}$, to an output context $\mathcal{V}$, where all are vectors. The context vector is deducted from $\mathcal{K}$ and $\mathcal{Q}$ which effectively calculates the context compatibility between $\mathcal{Q}$ with $\mathcal{K}$. Thus, the output of the 'Attention' mechanism is a mapping of $\mathcal{V}$ weighted by the compatibility of $\mathcal{K}$ with $\mathcal{Q}$. The $\mathcal{Q}$, $\mathcal{K}$ and $\mathcal{V}$ vectors can either come from the same sources (e.g., 'Self-Attention') or from different sources (e.g., 'Attention' in neural machine translation). In this model, they come from the same source and thus, 'Self-Attention' is used. For a give sequence of feature map $f_t$ $(t = 1 \dots T)$, the 'Attention' mechanism is described as in Vaswani et al. (2017):

$$Attention(\mathcal{Q}, \mathcal{K}, \mathcal{V}) = \text{Softmax}(\mathcal{Q}\mathcal{K}^{T_r})\mathcal{V} \tag{6.16}$$

Thus, the output of the 'Attention' mechanism is a mapping of Value $\mathcal{V}$ weighted by the compatibility of Key $\mathcal{K}$ with Query $\mathcal{Q}$. $T_r$ represents the transpose of a given vector/matrix. For 'Self-Attention' mechanism as is the case for this model, these three vectors are from the same source. The proposed 'Sequential Self-Attention' takes a query $f_t$ and maps against a set of keys $f_{t'}$ associated with the candidate feature maps from video frames at different time points in a given video. Then it return

values as an output context vector $\mathbf{v}_t$ which is computed by expanding Eq. 6.16:

$$\begin{aligned} \mathbf{v}_t &= \sum_{t'=1}^{T} a_{t,t'} f_{t'} \text{ and } a_{t,t'} = \text{softmax}(W_a g_{t,t'} + b_a) \\ g_{t,t'} &= \tanh(\mathcal{Q} + \mathcal{K} + b_g), \mathcal{Q} = \sigma(f_t W_g) \text{ and } \mathcal{K} = f_{t'} W_{g'} \end{aligned} \tag{6.17}$$

The above equation shows the decomposition of Eq. 6.16 to compute the queries $\mathcal{Q}$, the keys $\mathcal{K}$ and the values $\mathcal{V}$. The values are nothing but the output context vector $\mathbf{v}_t \in \mathcal{V}$. The weight matrices $W_g$ and $W'_g$ are for the respective feature maps $f_t$ and $f_{t'}$; $W_a$ is the weight matrix corresponding to their non-linear combinations. The element $a_{t,t'}$ is computed from $g_{t,t'}$ using the element-wise sigmoid function; $b_a$ and $b_g$ are the bias vectors. The 'Attention'-focused context vector $v_t$ conveys *how much to attend the feature map* $f_t$ in focus, *conditioned on its neighbourhood context* representing feature maps of all other video frames (see Figure 6.2). The weight matrices $W_g$ and $W'_g$, and the bias vectors $b_a$ and $b_g$ are learn-able parameters. The output context vector $\mathbf{v}_t$ is now fed into the next stage of the architecture, which is learn-able FV pooling.

### 6.4.4 Learn-able Fisher Vector Pooling

The output of the 'Sequential Self-Attention' is a sequence of context vectors $\mathbf{v} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_T\}$ corresponding to the input frames of the given video $v = \{F_1, F_2, \ldots, F_T\}$. The contextual information captures the neighbourhood context by considering all other surrounding frames. However, it does not capture the sequential information in a given sequence. The goal is to encode $\mathbf{v}$ using an internal state which summarises information extracted from the history of past observations. The internal state encodes the sequence knowledge and is responsible for making a decision on how to act. The widely used approach to model this internal state is through hidden units $h_t \in \mathbb{R}^n$ of a recurrent neural network and are updated over time. This is achieved in the next step by using a fully-gated Bi-LSTM. The Bi-LSTM is a concatenation of two LSTMs in which one is focused on the forward direction (i.e., $\mathbf{v}_1 \ldots \mathbf{v}_T$) and the other one is on the backward direction (i.e., $\mathbf{v}_T \ldots \mathbf{v}_1$). Figure 6.3, presents an unrolled Bi-LSTM for a better understanding of the temporal dependency, but in reality it is the same Bi-LSTM. The Bi-LSTM generates output as a sequence of hidden states in forward direction $h = \{h_1, h_2, \ldots, h_T\}$ and backward direction $h^{'} = \{h^{'}_1, h^{'}_2, \ldots, h^{'}_T\}$ corresponding to the input sequence of context vectors $\mathbf{v} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_T\}$. The hidden states in both forward and backward direction are concatenated $\mathbf{h} = [h, h^{'}]$ to produce the final contextual feature vector for further processing.

Generally, the sequence recognition using a LSTM is carried out by considering the associated features at the last time step $T$ and the previously involved hidden states. This is a fundamental flaw in LSTM since it uses recurrent connections to maintain and communicate temporal information. Thus, researchers have recently explored dynamical temporal pooling (Yeung et al., 2018) as an additional direct pathway for referencing previously seen frames. Inspired by this approach, this method focuses on the hidden states of the Bi-LSTM and let the model *learns to attend* the different

parts of the hidden states $h$ and $h'$ at each step of the output generation. This is achieved by using learn-able pooling with FVs in which the similar hidden states of the Bi-LSTM are grouped together via clustering. Instead of calculating the FVs from GMM as described in the last section, NetFV (FV integrated with a neural network) is used to learns these parameters (Miech et al., 2017). The main idea is to assign $\mathbf{h}_t$ to the cluster $k$ as a soft assignment which is done by changing the Eq. 6.12 to the following:

$$\alpha_k(\mathbf{h_t}) = \frac{e^{W_k^{Tr}\mathbf{h_t}+b_k}}{\sum_{j=1}^{K} e^{W_j^{Tr}\mathbf{h_t}+b_j}} \tag{6.18}$$

Here the weight $W_j$ and the bias vector $b_j$ are learn-able parameters. The soft assignment of $\alpha_k(\mathbf{h_t})$ of hidden state $\mathbf{h_t}$ to cluster $k$ measures how close the hidden state $\mathbf{h_t}$ is to cluster $k$. As in the previous section, $j \in (1, K)$ where $K$ is the total number of clusters. This is different from the original FV (Perronnin; Dance, 2007) in the sense that the cluster centres $c_k$ and the co-variance matrices $\sigma_k$ are not coupled to the cluster weights $\alpha_k$. Using the above soft assignment, the FV is computed using the NetFV representation by Miech et al. (2017), which is adapted from Eqs. 6.13 and 6.14 respectively:

$$\begin{aligned} FV_1(j,k) &= \sum_{t=1}^{T} \alpha_k(\mathbf{h_t}) \left( \frac{\mathbf{h_t}(j) - c_k(j)}{\sigma_k(j)} \right) \\ FV_2(j,k) &= \sum_{t=1}^{T} \alpha_k(\mathbf{h_t}) \left( \left( \frac{\mathbf{h_t}(j) - c_k(j)}{\sigma_k(j)} \right)^2 - 1 \right) \end{aligned} \tag{6.19}$$

FV $FV_1$ and $FV_2$ capture the respective first-order and second-order statistics. As in the last section, $c_k$ and $\sigma_k$ are the learn-able cluster centers and diagonal co-variances of the $k_{th}$ clusters, where $k \in [1, K]$. Moreover, $c_k$ and $\sigma_k$ are learned independently from the parameters of the soft assignment $\alpha_k$ as in Eq. 6.18. In adapting the above equations from Eq. 6.13 and Eq. 6.14 the normalising factor $w_k^{1/2}$ is left out as both $FV_1$ and $FV_2$ are L2 normalised later. The FVs are then concatenated to get the final $FV = [FV_1, FV_2]$. The current implementation is different from the approach of Miech et al. (2017), since weighted pooling mechanism is used in an activity-aware manner and is defined as:

$$Pooling(FV) = \text{softmax}(W_p FV + b_p) \tag{6.20}$$

where matrix $W_p \in \mathbf{R}^{|FV| \times C}$ and bias vector $b_p$ are learn-able parameters and $C$ is the number of human activity classes. Therefore, by integrating FV to the neural network model to exploit the hidden Bi-LSTM states, this work presents a novel end-to-end trainable model. The next section describes the experiments carried out to evaluate the model performance.

## 6.5 Experiments, Results and Analysis

### 6.5.1 Implementation

In order to train the model, frames are uniformly sampled from each video. Sub-sampling is done with uniformly sampled 20 and 30 frames from each video from the MSR and NTU datasets respectively. The default frame size of $299 \times 299$ is used as the input to Inception ResNet-V2 (Szegedy et al., 2017) model. Data augmentation including random scaling (factors ranges from 0.75 to 1.25) and rotation (angle $\pm 15$ degrees) is applied to each set of video sequences to improve the generalisability of the proposed model. The model is implemented using Tensorflow framework with Keras wrapper. For transfer learning, the pre-trained CNN model is fine-tuned for 5 epochs. The whole model is trained for further 25 epochs. Adam optimiser (Kingma; Ba, 2014) is used to optimise the categorical cross-entropy $E_v$ in Eq. 6.15 with an initial learning rate of 1e-4, and a learning rate decay of 1e-6. The model is trained with a mini-batch size of 4 to fit with a GPU memory of 24 GB. To train the model, a Linux PC (Ubuntu 16.04 LTS) with a Nvidia Quadro P-6000 GPU has been used.

### 6.5.2 Experiments and Results

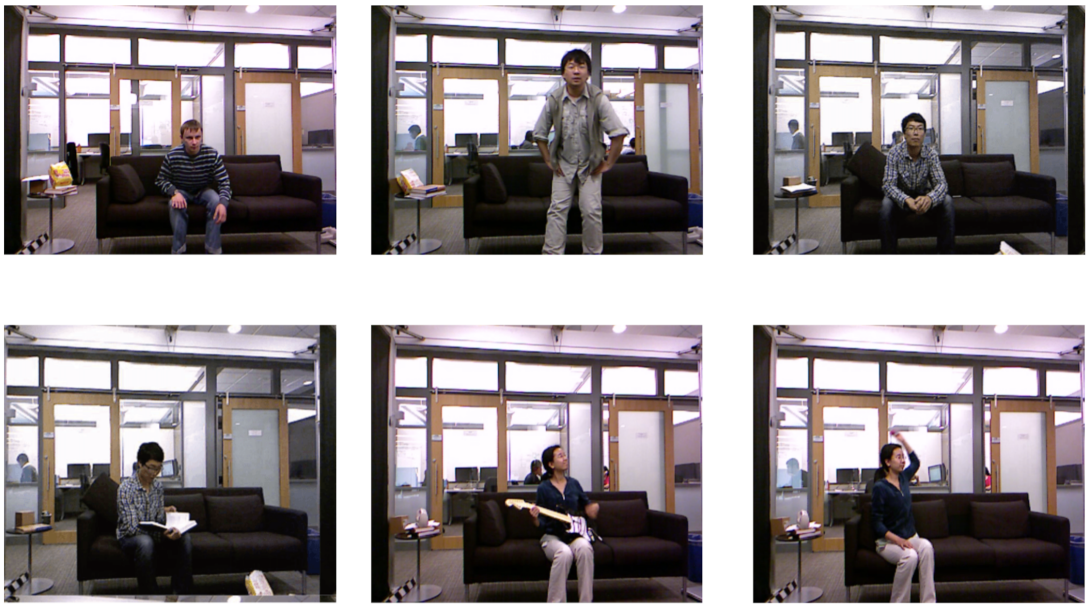


Figure 6.4: Sample of the MSR dataset (Wang et al., 2012). Clockwise from top-left: standing up, sitting down, sitting, throwing, playing guitar, reading

The model was evaluated on two challenging daily activity recognition datasets. The first one is the MSR 3D Daily activity dataset (Wang et al., 2012). This dataset contains a total of 320 videos containing 16 different daily activities such as drinking, eating, etc. There are 10 subjects who perform each of the 16 activities twice resulting in 320 video sequences. For evaluation, the standard protocol of Wang et al. (2012) is followed in which 50% of the subjects, i.e., subjects 1 to 5 are used for training and the remaining 50% is used for evaluation. The evaluation protocol is challenging since only half of the data is used for training the model.

The second is the NTU RGBD dataset (Shahroudy et al., 2016), which is one of the largest human

Figure 6.5: Sample of the NTU-RGBD dataset (Shahroudy et al., 2016). Clockwise from top-left: drinking, eating, dropping, dropping, standing, sitting, kicking, pushing, hugging, cross hand at front, taking a selfie, phone call

activity recognition datasets. This dataset contains approximately 57K video samples of daily activities. There are 60 different activity classes performed by 40 different subjects. The activity classes include person-objects interaction, single-person activities (e.g., jumping, waving hands, etc.) and person to person interactions like handshaking. The authors specify cross-subject and cross-view protocols for evaluation. In this study, the cross-subject protocol is used, which is more difficult than the cross-view protocol.

| **Methods** | **Pose** | **RGB** | **Accuracy** (%) |
|---|---|---|---|
| Ensemble (Wang et al., 2012) | × | - | 68.0 |
| Efficient Pose (Eweiwi et al., 2014) | × | - | 73.1 |
| Moving Pose (Zanfir et al., 2013) | × | - | 73.8 |
| Poselets (Tao; Vidal, 2015) | × | - | 74.5 |
| PDA (Baradel et al., 2018a) | - | × | 75.3 |
| Actionlet (Wang; Wu, 2013) | × | - | 88.8 |
| PDA (Baradel et al., 2018a) | × | × | 90.0 |
| **Proposed Approach** | - | × | **91.9** |

Table 6.1: Comparison of the proposed model with the state-of-the-art approaches on MSR 3D daily activity dataset (Wang et al., 2012)

The standard top-1 accuracy was used as evaluation metric. The performance of the proposed model and state-of-the-art approaches using the MSR Activity dataset is presented in Table 6.1. It is clear that the proposed approach (91.9%) outperforms the state-of-the-art approaches by a significant

margin. For example, using only the RGB video, the current approach is 1.9% higher than the approach of Baradel et al. (2018b) (90%) which combines multi-modal information (pose and RGB). Using only RGB information, the accuracy (75.3%) of Baradel et al. (2018b) is significantly inferior to the proposed approach (91.9%). This suggests the benefit of the proposed 'Attentional Learn-able Pooling' mechanism for human activity recognition using only RGB information. In Table 6.1, most of the state-of-the-art approaches are based on the body pose represented as a 3D skeleton. The performance of the current approach in which only RGB data is used is better than the existing approaches (Table 6.1). This justifies that the proposed approach is easily applicable to video-based activity recognition without requiring additional information such as depth, which is essential for the computation of 3D skeletons.

| **Methods** | **Pose** | **RGB** | **Accuracy** (%) |
|---|---|---|---|
| Part-aware LSTM (Shahroudy et al., 2016) | × | - | 62.9 |
| C3D (Tran et al., 2015) | - | × | 63.5 |
| DSSCA-SSLM (Shahroudy et al., 2017) | × | × | 74.9 |
| Synthesised CNN (Liu et al., 2017) | × | - | 80.0 |
| ST-GCN (Yan et al., 2018) | × | - | 81.5 |
| DPRL+GCNN (Tang et al., 2018) | × | - | 83.5 |
| PDA (Baradel et al., 2018a) | × | × | 84.8 |
| 3-Scale ResNet152 (Li et al., 2017a) | × | - | 85.5 |
| Glimpse Clouds (Baradel et al., 2018b) | - | × | 86.6 |
| **Proposed Approach** | - | × | **87.2** |

Table 6.2: Performance of the proposed model in comparison to the state-of-the-art approaches on NTU RGBD dataset (Shahroudy et al., 2016). All the results are in cross subject settings which is more challenging than the cross view settings

Table 6.2 presents the performances of the proposed approach and state-of-the-art approaches using the NTU dataset (Shahroudy et al., 2016). Similar to the performance in MSR Activity dataset, the proposed approach (87.2%) outperforms the state-of-the-art approaches in which many of them use multi-modal information (RGB + Pose). Using RGB only, the proposed approach is 0.6% better than the best performing approach (Glimpse Clouds (Baradel et al., 2018b)) and 23.7% better than the approach of Tran et al. (2015) that uses only RGB data. It is also clear that the proposed approach is significantly better than the 3D skeleton-based approaches. This signifies the proposed 'Attentional Learn-able Pooling' mechanism plays a key role in discriminating human activities in videos.

### 6.5.3 Ablation Study

| Base CNN Network | Params | Base Acc | Proposed Acc |
|---|---|---|---|
| MobileNets (Howard et al., 2017) | ∼4.2M | 75.0% | 79.4% |
| NasNet Mobile (Zoph et al., 2018) | ∼2.6M | 79.0% | 82.5% |
| Inception V3 (Szegedy et al., 2016) | ∼23M | 79.5% | 84.0% |
| Inception ResNet-V2 (Szegedy et al., 2017) | ∼54M | **86.9**% | **91.9**% |

Table 6.3: Comparison of base network accuracy on the MSR dataset (Wang et al., 2012). 'Base Acc' implies the performance of the core CNN-LSTM models without the use of the proposed Sequential 'Self-Attention' and novel learn-able pooling using FV. The associated parameters are presented as the nearest millions

In this section, three different experiments have been conducted to justify the suitability of various components integrated to the current model. These are: 1) different state-of-the-art deep CNN models to extract CNN features for the network, 2) the benefits of the proposed 'Sequential Self-Attention' in comparison to the 'Multi-Head Attention' and 3) compare the performance of the proposed FV-based learn-able pooling with the traditional GAP and FC combination. First, the performance using different base CNNs to extract frame-wise CNN features from videos is analysed. Here, three state-of-the-art CNN models with different characteristics are used. The performance on MSR dataset (Wang et al., 2012) is shown in Table 6.3. For the base network, the last layer (i.e., classification) is comprised of a GAP layer followed by a FC layer with Soft-max activation. This is placed on top of the core CNN-LSTM network. The NasNet Mobile (Zoph et al., 2018) outperforms the MobileNets (Howard et al., 2017). It also has significantly fewer parameters (∼2.6M vs ∼4.2M) in comparison to the MobileNets. Among the three architectures, the Inception-ResNet-V2 (Szegedy et al., 2017) achieves the best accuracy and has the largest number of parameters (∼54M). The proposed algorithm also improves accuracy when Inception-V3 (Szegedy et al., 2016) as a backbone. Although the proposed model benefits from a better backbone (Inception-ResNet-V2), it is able to improve results across 4 different backbones. From Table 6.3, it is evident that the performance of the proposed approach with the novel 'Sequential Self-Attention' and learn-able FV pooling on top of the core CNN-LSTM network is significantly better than the base accuracy. This demonstrates the applicability of the proposed method across a spectrum of CNNs ranging from lightweight to heavier models.

| Dataset | Base Acc | MHA Params | SSA Params | MHA Acc | SSA Acc |
|---|---|---|---|---|---|
| MSR (Wang et al., 2012) | 86.9% | ∼9.4M | ∼98K | 90.6% | **91.3**% |
| NTU (Shahroudy et al., 2016) | 82.2% | ∼9.4M | ∼98K | 86.3% | **86.6**% |

Table 6.4: Comparison of the performance of the proposed 'Sequential Self-Attention' (SSA) with the 'Multi-Head Attention' (MHA). The classification layer consists of the combination of GAP and FC

Second, the effectiveness of the proposed 'Sequential Self-Attention' in comparison with the 'Multi-Head Attention mechanism' (Vaswani et al., 2017) is demonstrated. The 'Multi-Head Attention' mechanism focuses on more important parts of the feature map in discriminating various activities. The results are shown in Table 6.4, using both the MSR Activity (Wang et al., 2012) and NTU-RGBD (Shahroudy et al., 2016) datasets. The performance of both 'Attention' mechanisms significantly improves the recognition accuracy in comparison to the base accuracy. In case of 'Multi-Head Attention' mechanism (Vaswani et al., 2017), the keys $\mathcal{K}$, queries $\mathcal{Q}$ and values $\mathcal{V}$ vectors are transformed through a number of trainable weights. Each transformation produces a different mapping of the same input vectors. Each mapping is called 'head' and hence the name 'Multi-Head Attention'. The optimum number of heads for 'Multi-Head Attention' is 4 and is found experimentally. The performance of the proposed 'Sequential Self-Attention' is better than the 'Multi-Head Attention'. Moreover, the associated number of learn-able parameters with 'Sequential Self-Attention' (∼98K) is significantly less than the 'Multi-Head Attention' (∼9.4M). It can be assumed that owing to more parameters the model may be over-fitting with 'Multi-Head Attention' mechanism.

| Dataset | Base | SSA & GAP/FC | SSA & FV pooling |
|---|---|---|---|
| MSR (Wang et al., 2012) | 86.9% | 91.3% | **91.9**% |
| NTU (Shahroudy et al., 2016) | 82.2% | 86.6% | **87.2**% |

Table 6.5: Impact of Sequential 'Self-Attention' and the novel FV pooling. The base network is Incpetion-Resnet-V2 + LSTM + GAP/FC

In the third, the impact of the novel learn-able activity-aware pooling using FV on model's recognition performance is studied. In this experiment, the recognition accuracy of the proposed model is compared using the proposed AAP with the customary combination of the GAP and FC layer. The performance is presented in Table 6.5 using both the MSR Activity (Wang et al., 2012) and NTU-RGBD (Shahroudy et al., 2016) datasets. It is evident that the recognition accuracy is significantly better when the FV-based activity-aware pooling mechanism is used. This is mainly because the proposed mechanism learns semantic clusters from hidden Bi-LSTM states. The enables more effective pooling to represent a high-level encoding of the spatio-temporal structure in videos and thus, achieve better performance. The number of clusters is a tune-able hyper-parameter and the optimal number of clusters is experimentally found. These values are 32 and 64 for the MSR Daily Activity (Wang et al., 2012) and NTU-RGBD dataset (Shahroudy et al., 2016), respectively. Figure

6.6 presents the confusion matrix of the model for MSR dataset while Figure 6.7 represents the same for the NTU-RGBD dataset.

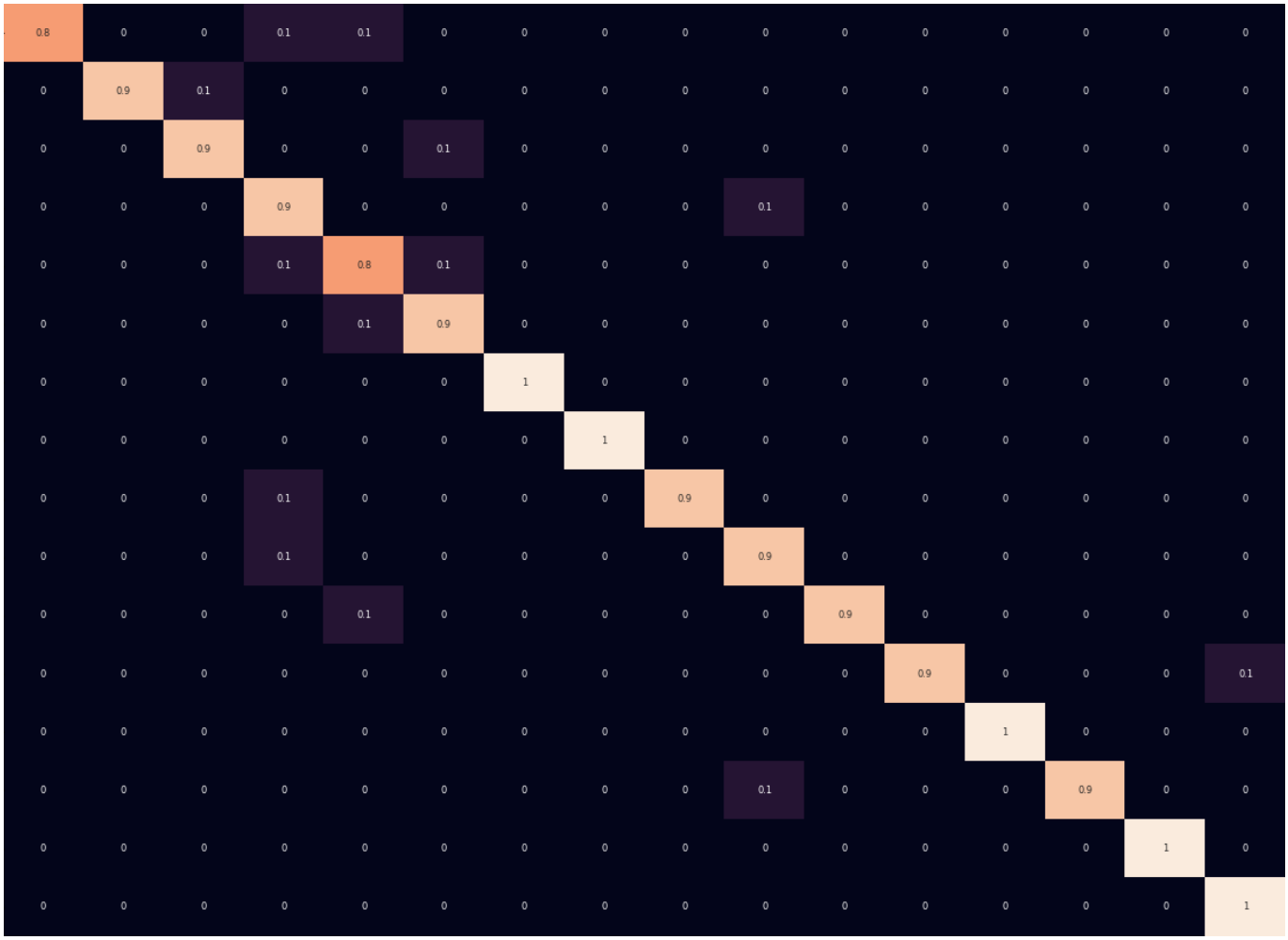



Figure 6.6: Confusion Matrix of the monocular video-based classifier with an accuracy of 91.9% (Chapter 6, Table 6.1) for the MSR dataset

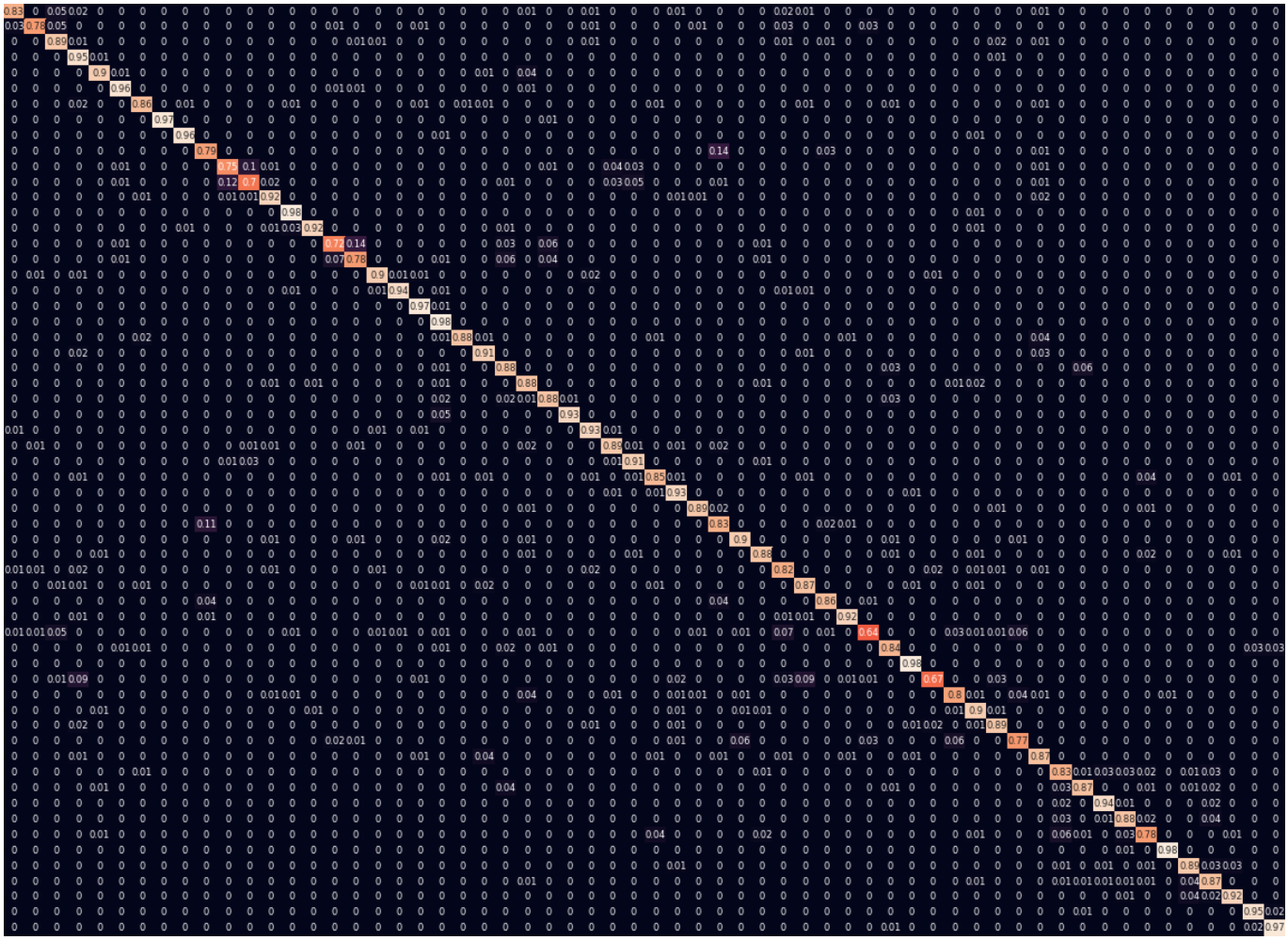


Figure 6.7: Confusion Matrix of the monocular video-based classifier with an accuracy of 87.2% (Chapter 6, Table 6.2) for the MSR dataset

## 6.6 Discussion

This Chapter introduces a simple yet effective approach to recognise human activities using only monocular RGB videos. The proposed novel 'Attention'-based learn-able pooling mechanism can be easily integrated into any of the existing deep CNN models used for image/object or activity recognition. The method involves a 'Sequential Self-Attention' mechanism to capture the contextual information which conveys how much to attend (give weightage) to a feature map conditioned on its neighbouring feature maps. Further, an alternative to the customary GAP and FC layer is presented with a learn-able pooling mechanism in the form of the FV that uses learn-able semantic clusters to capture the first-order and second-order statistics. This novel activity-aware pooling mechanism learns structural information from hidden states of a bidirectional LSTM to provide more effective structure learning and representation. The model is evaluated using two challenging datasets and preforms better than the state-of-the-art. The literature review (Chapter 3 (Sec. 3.2.3) showed that authors have researched for better alternatives to statistical pooling mechanism with learn-able pooling mechanisms. This proposed novel FV-based learn-able pooling is a contribution in this direction. Further analysis of the FV-based activity-aware pooling mechanism is presented in Chapter 8.

With regards to the current study, the Chapter caters to the fourth objective. The review on CV-based rehabilitation and assessment (Chapter 2.7, Sec. 2.7) shows that one of the several assessment approaches is human activity recognition where authors have used their own dataset. However, to prove efficacy of any proposed model and to be accepted in peer-reviewed conferences/journals, it is necessary to use well-known benchmark datasets for evaluation. In the absence of publicly available, large-scale datasets in the domain of CV-based assessment and rehabilitation, two well-known datasets have been used that present normal ADL. Although the method presented in this Chapter is applied to human activity recognition using monocular RGB videos, the idea is also applicable to other sequence recognition problems. The FV-based learn-able pooling mechanism has been used to improve the performance of the purely pose-based multi-label activity recognition model presented in Chapter 8. This research has been accepted for presentation at the *IEEE ICRA 2021*.

## 6.7 Conclusion

This Chapter introduces a novel approach for human activity recognition using monocular RGB videos. However, with 3D human pose-based datasets easily available, researchers have exploited the combination of joint position and RGB images to achieve state-of-the-art results. The next Chapter presents another approach for activity recognition which combines RGB and 3D pose information. It is a fact that RGB video requires extensive computational power. The models presented in this Chapter and the next have huge number of parameters (54 million) owing to the high-performing Inception-ResNet-V2 CNN architecture used as base. With 3D pose information readily available, many authors are now turning towards exclusive pose-based models. Thus, in the multi-label activity recognition (Chapter 8), a lightweight purely pose-based model is explored.

# Chapter 7

# Human Activity Recognition: Model 2

## 7.1 Introduction

This Chapter presents a second human activity recognition model which corresponds to the fourth objective of this project. The previous Chapter (Chapter 6, Sec. 6.1) explains that the study approaches the problem of multi-label ('Activity' and 'Impairment') ADL recognition by first focusing on single-label ('Activity') ADL recognition. Accordingly, an ADL recognition model based on monocular RGB videos was presented in the previous Chapter. The current Chapter continues exploring ADL recognition. But, in contrast to the previous Chapter, explores a model that combines RGB video data with 3D pose information. The processing of RGB video data is done in a manner similar to the previous model (Chapter 6). This Chapter proposes a novel method to encode the human body pose information in the pose network. The encoding exploits the structural relationships and dependencies between various body joints, as well as captures long-term temporal dependencies of each body joint. Generally, in pose-based models, each joint is represented by a vector of length three consisting 3D joint positions (Shahroudy et al., 2016). Sometimes, this is enriched with other information such as distance between two body joints, pairwise relations which presents an augmented and enriched vector for each joint (Vemulapalli et al., 2014). Instead of encoding additional handcrafted pose-related features, the proposed 'Attention'-driven body pose network learns such encodings. Combining this information with RGB data influences the network to focus on important spatial points in the RGB data while capturing other contextual cues such as an object in hand and so on. The novel 'Attention'-based multi-stream deep architecture presented in this Chapter, introduces learn-able encodings from 3D pose data that outperforms the state-of-the-art approaches on three challenging datasets.

The next section describes the rationale behind the approach, followed by a detailed description of proposed approach. This is followed by a section on experiments and analysis where the experiments performed are described and the results are compared with existing state-of-the-art approaches. The analysis includes an ablation study which highlights the impact of each part of the model on the overall performance. Finally, a discussion is presented which highlights the contribution of this model with regards to the rest of the study and the broader research area.

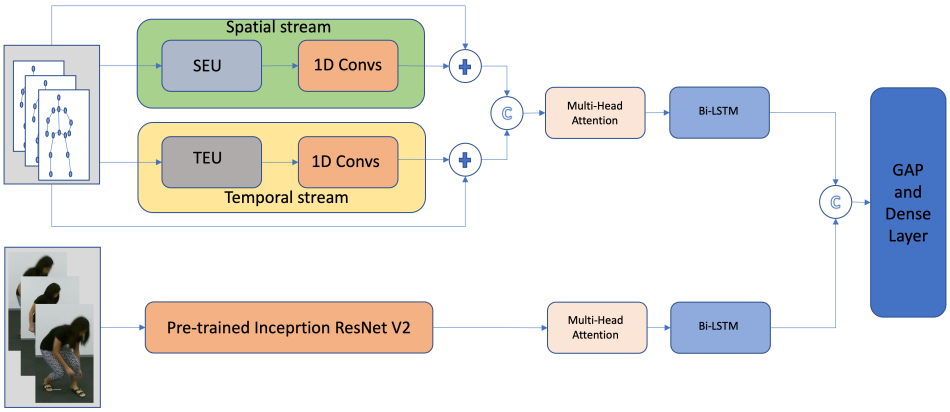



Figure 7.1: A novel skeleton sequence encoding approach is introduced through learned joint encodings. The SEU learns the structural dependencies and relationships between various body joints and presents a spatially enhanced sequence to the network. The TEU learns the frame-wise position of each joint to learn a temporally augmented meaningful representation. Both the streams are processed through 'Multi-Head Attention' mechanism (Vaswani et al., 2017). The '+' symbol stands for addition while 'C' indicates concatenation

## 7.2 Motivation/Rationale

The previous Chapter explains the importance of recognising human activities from RGB video data. However, with the availability of cheap commercial devices such as Kinect, both RGB videos and human body skeleton (represented by 3D body joints) are readily available. Moreover, the availability of large-scale datasets (Shahroudy et al., 2016) with both RGB-D and skeleton information has significantly contributed to advancing the field. RGB data provides various visual, temporal and contextual cues linking a given human activity. Moreover, RGB videos contain detailed information regarding the scene as well as objects handled by subjects, which can provide contextual information vital in discriminating various human activities. On the other hand, pose information contains 3D positions of each joint for each frame and can be processed as sequential information for human activity recognition (Kim; Reiter, 2017). Authors (Baradel et al., 2018a; Shahroudy et al., 2017; Baradel et al., 2018b) combining pose and video data have benefited from information contained within both the data modalities. Thus, this chapter explores a combined video-pose model that benefits from both the modalities. There are many aspects of designing and developing an efficient model that combines RGB and pose information. The proposed model aims to address the following aspects to impact the model performance positively:

1. Video sequences contain a high amount of visual, as well as temporal information. Deep CNN models are very good at capturing visual (spatial) information, but they lack the ability to semantically process temporal information.

2. Pose-based models should also be able to learn the spatial relationships between various body-joints in order to semantically encode the structural relationships and various inter-dependencies among different body parts.

Modelling the above-mentioned main points and combining the multi-modal information in a meaningful way was the key to overcome some of the challenges in human activity recognition problem. The proposed model aimed to address this by developing an end-to-end deep architecture consisting of

two pose and one RGB stream as shown in Figure 7.1. A pre-trained Inception-ResNet-V2 (Szegedy et al., 2017) is used to process the RGB video in a manner similar to the model presented in the previous Chapter (Chapter 6). This is followed by 'Self Multi-Head Attention' (Vaswani et al., 2017) mechanism and a Bi-LSTM. While the pre-trained network effectively captures the spatial information, the Bi-LSTM learns to capture the temporal information in videos. The 'Multi-Head Attention' mechanism further enhances the CNN-LSTM network performance by focusing the network on the spatial features that are important for discrimination.

The skeleton information is processed through a spatial stream and a temporal stream in parallel. The streams are then concatenated and passed through a 'Self Multi-Head Attention' mechanism followed by a Bi-LSTM. The spatial stream in the pose network consists of a SEU and similarly the temporal stream includes a TEU. The SEU provides an enriched representation that *learns to capture* the structural relationship between various body joints at each frame in a given sequence. This presents a spatially enhanced representation of the skeleton sequence to the network that is *learned*. On the other hand, the TEU encodes the temporal relationship of each body joint over the duration of a given sequence to present a temporally enriched representation of the pose sequence.

## 7.3 Proposed Approach: ADL Recognition Model 2

The architecture of the proposed network is shown in Figure 7.1. The model takes as input a video sequence and the corresponding body pose sequence and provides output as 'Activity' class label to the input sequence. The model introduces a novel two-stream 'Attention'-based joint position encoding framework that temporally and spatially learns the structural relationships between various body joints. The feature maps describing the spatial and temporal structures are concatenated in the final representation. Afterwards, the concatenated skeleton stream is combined with an 'Attention'-based time distributed CNN network in a late fusion mode (Figure 7.1). From the literature review (Chapter 7, Sec: 3.4.2), it is observed that instead of presenting sequences of joints directly to the network, many state-of-the-art approaches (Demisse et al., 2018; Zanfir et al., 2013; Ke et al., 2017; Vemulapalli et al., 2014; Wang et al., 2012) have tried to learn more enriched representations. For example, grouping of joints through a hierarchical network (Wang et al., 2012), enriching the representations by presentation of hand-crafted features (Vemulapalli et al., 2014). In contrast, the proposed model automatically learns such enhanced representations through the SEU and the TEU. Generally, a sequential network such as TCN or RNN only learns the temporal relationship between frames. On the other hand, the proposed network i) learns the structural relationships between various body joints, and ii) learns the frame-wise relationship of each joint, in addition to learning the temporal relationships between frames. In the following subsections, the SEU, the TEU, the RGB stream and the 'Attention' mechanism used in the model are elaborated.

### 7.3.1 Pose Network: Spatial Stream

The spatial stream consists of a SEU followed by three layers of 1D convolutions. The goal of the SEU is to present an enriched pose information to the network for improved performance. Normally,

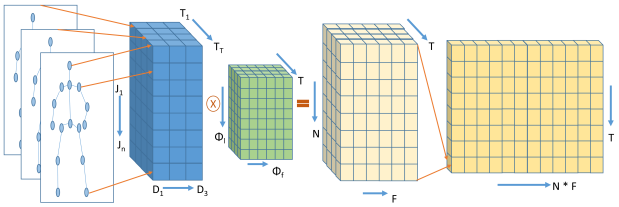


Figure 7.2: The SEU augments the spatial data with learned representations. Typically, a matrix of size $T, J * D$ is presented for sequential processing; instead, a learned representation of size $T, N * F$ is presented. $T$ is time or number of frames, $F$ is the number of filters and $J$ is the number of joints. $D$ is normally 3 representing 3D positions and is often enhanced with additional hand-crafted features including, but not limited to velocity and acceleration. Instead, the SEU learns $F$ representations per $J$ joints per $T$ time-steps. The 'X' symbol indicates convolution

for sequential processing, the input consists of 2D or 3D joint coordinates for each frame. To present a richer representation, a three-layer 1D convolutional network is used which learns the structural information between various body joints. Let there be $T$ number of frames in a sequence, $J$ is the number of body joints and $F$ is the total number of filters in a layer. A 1D convolution operation performs the following mapping with input vector V $\in \mathbf{R}^{F \times JD}$:

$$M_{T,F} = U(\Theta_F, V_{T,J*D}) \tag{7.1}$$

where $U$ is the convolution operation parametrised by filters $\Theta_F$ and $D$ indicates the number of dimensions of each body joint which (in this case is 3 for 3D pose). Instead of performing the convolution operation on the entire input map, the SEU performs the convolutions for each frame separately. Thus, for each frame $t \in \{1 \dots T\}$ in the sequence, pose vector $V \in \mathbf{R}^{J \times D}$ is encoded through 1D convolution operations. Formally, the following operation is carried out:

$$M^t_{J,F} = U_t(\Theta_F, V_{J,D}) \tag{7.2}$$

$$M^t_{J,F} \rightarrow M_{T,J*F} \tag{7.3}$$

As shown in Eq. 7.2, for each time step or frame, a convolution operation is performed where each joint is represented individually. The learned map is then spatially squeezed and aggregated temporally as shown in Figure 7.2 (Eq. 7.3). Normally, while encoding skeleton sequence, a 2D vector of dimensions $(T, J * 3)$ is presented to the network. Instead, a learned representation of size $(T, J * F)$ is presented. For each joint $j$, in every frame at position $t \in \{1 \dots T\}$, there are $F$ filters representing the learned encoding. The extra $F - D$ representations per joint captures the spatial relationships between various body joints for each frame. Normally, the skeleton sequence is enriched (augmented) with hand-crafted mechanisms such as groups of joints (Wang et al., 2012), velocity and acceleration (Zanfir et al., 2013) and so on. In contrast, the proposed model is able to *'learn'* such representations. This enriched representation is presented to the spatial stream which consists of 3 layers of 1D convolutions.

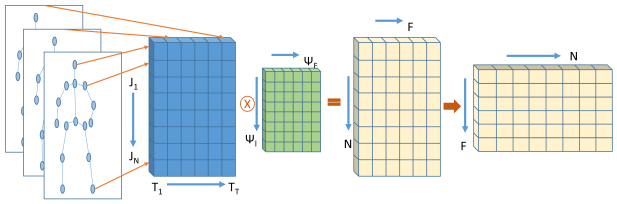



Figure 7.3: The TEU encodes frame-wise positions of each individual body joint to learn temporally augmented representations. Instead of temporal length $T$, learned temporal sequence of length $F$ determined by the number of filters is presented. The 'X' symbol indicates convolution

### 7.3.2 Pose Network: Temporal Stream

Similar to the spatial stream, the temporal stream consists of the TEU followed by three layers of 1D convolutions. The goal of the TEU is to encode the frame-wise positions of body joints and present a temporally enhanced representation for each joint, individually. Similar to the SEU, three layers of 1D convolutions is used. For each joint $j \in \{1 \dots J\}$, a vector is encoded $\bar{V} \in \mathbf{R}^{J \times T}$, through 1D convolutions parameterised by filters $\Psi_F$. Formally, the following transformation is performed:

$$\bar{M}_{J,F} = \bar{U}(\Psi_F, \bar{V}_{J,T}) \tag{7.4}$$

$$\bar{M}_{J,F} \rightarrow \bar{M}_{F,J} \tag{7.5}$$

In contrast to a normal convolution operation (Eq. 7.1), the TEU outputs a feature map of dimensions $(F, J)$ instead of $(T, J * F)$. Thus, as shown in Figure 7.3, the TEU represents a map with temporal size $F$ instead of $T$. This is equivalent to augmenting the temporal dimension of the input vector from $T$ to $F$ based on the number of filters. Enhancing the temporal dimension in such a manner benefits the network. The enhanced temporal representation is fed into the temporal stream, which consists of three layers of 1D convolutions.

### 7.3.3 Pose Network: Stream Fusion

The impact of residual connection is well-studied (He et al., 2016) and the proposed network also benefits from residual connections. The resulting maps from both the temporal and spatial streams are added to residual connections as shown in Figure 7.1. Ba et al. (2016) argue that for sequential networks, layer normalisation is beneficial when compared to batch normalisation. It is observed that the model also benefits from the layer normalisation, which is added to the residual connections. The streams are then concatenated along the time axis. As a result, the fused pose stream has a compact yet richer representation of both the spatial and the temporal characteristics of the given sequence. In order to exploit this representation further, 'Multi-Head Attention' with Bi-LSTM is used as shown in Figure 7.1. The 'Attention' mechanism learns weighted representations of different temporal sub-spaces to focus the network on important temporal zones for discrimination. The representations of the 'Attention'-based fused pose stream is very rich and is different from the individual stream. Therefore, it is beneficial to exploit the same with further processing based on Bi-LSTM to capture the long-term sequential information.

### 7.3.4 RGB Stream: Context/scene Descriptor

Unlike body-pose, video frames provide richer information which is explored to describe the contextual/scene descriptor. As shown in Figure 7.1, a pre-trained Inception-Resnet-V2 (Szegedy et al., 2017) model is used to extract this contextual scene descriptor. The details of the RGB stream is same as the model presented in the previous Chapter and is extensively elaborated in the previous Chapter (Chapter 6, Sec. 6.4.2). The only difference is that the output of the CNN is fed into 'Sequential Self-Attention' mechanism in the previous model (Chapter 6, Sec 6.4.3), whereas here it is fed into a 'Self Multi-Head Attention' mechanism. The reason for using 'Self Multi-Head Attention' mechanism is that 'Sequential Self-Attention' did not work well when the RGB stream was concatenated with the pose-data. Vaswani et al. (2017) propose the of 'Multi-Head Attention' mechanism as an improvement on 'Sequential Self-Attention'. Therefore, it is reasonable to expect better performance from 'Multi-Head Attention' mechanism. However, as explained in the previous Chapter (Sec. 6.5.3), over-fitting may have contributed to 'Multi-Head Attention' performing less than 'Sequential-Self Attention'. Sharma et al. (2016) observe that weighting the 3D CNN outputs through 'Attention' mechanism provides higher recognition accuracy. Inspired by this approach, 'Multi-Head Attention' mechanism (Vaswani et al., 2017) from machine translation problem is adapted to map the output of the Inception-Resnet-V2 CNN model to a weighted version of itself. Instead of using 3D CNN outputs as in (Sharma et al., 2016), the proposed model uses average pooled 2D feature maps from the Inception-Resnet-V2 model. As a result, it reduces the network size and parameters. Experimentally it was found that the model benefits from 'Multi-Head Attention'-based temporal processing. The details of the 'Self Multi-Head Attention' mechanism is described in the next section.

### 7.3.5 Attention Mechanism

In general, all 'Attention' mechanisms maps input values $\mathcal{V}$ to weighted representations using keys $\mathcal{K}$ queries $\mathcal{Q}$. As a result, values $\mathcal{V}$ focuses on more discriminatory features. In machine translation, where encoder-decoder style architectures are normally used, $\mathcal{K}$ and $\mathcal{V}$ are obtained from decoder while $\mathcal{Q}$ is from the encoder. For Self-Attention mechanisms which essentially calculates weighted representations of itself, $\mathcal{K}$, $\mathcal{Q}$ and $\mathcal{V}$ is the same vector. In case of 'Multi-Head Attention' (Vaswani et al., 2017), the input vector is divided into a number of parts called heads. 'Attention' mapping is carried out for each head separately and the heads are linearly concatenated in a weighted manner to keep the input dimension same as the output. This results in the output maps focusing on different sub-spaces of the input vector. Formally, 'Multi-Head Attention' (Vaswani et al., 2017) can be represented as:

$$\text{Attention}(\mathcal{Q}, \mathcal{K}, \mathcal{V}) = \text{softmax}(\mathcal{Q}\mathcal{K}^{T_r}/\sqrt{d_k})\mathcal{V} \tag{7.6}$$

$$\text{MultiHead}(\mathcal{Q}, \mathcal{K}, \mathcal{V}) = \text{Concat}(head_1, \ldots, head_h)W^o \tag{7.7}$$

$$\text{where } head_i = \text{Attention}(\mathcal{Q}W_i^Q, \mathcal{K}W_i^K, \mathcal{V}W_i^V) \tag{7.8}$$

where $T_r$ represents the transpose of a given vector/matrix. In this case, the mechanism is applied in a Self-Attention manner which means $\mathcal{K}, \mathcal{Q}, \mathcal{V}$ are from the same input vector. As a result, $W_i^Q = W_i^K = W_i^V \in \mathbf{R}^{D \times D}$ and $W^o \in \mathbf{R}^{hD \times D}$. When this is applied to the RGB stream, the final dimension is $D = 1536$, which is same as the output of the Inception-Resnet-v2 (Szegedy et al., 2017) network. For pose network, $D = 120$ and is the enriched feature length obtained from the concatenation of SEU and TEU. The number of 'Attention' heads is experimentally found to be four, i.e., $h = 4$. It is also shown that the adapted 'Multi-Head Attention' mechanism increases the recognition accuracy for both the RGB and the fused posed streams. Here, $d_k = D/h$ is a scaling factor. After the 'Attention' mechanism, a Bi-LSTM is used in a manner similar to the previous model, details of which are mentioned in the previous Chapter (Chapter 6, Sec. 6.4.4).

### 7.3.6 Combined Model: Fusion of three streams

Multiple streams in a given model are usually combined using either early fusion or late fusion. In case of the proposed model, a hybrid approach is used, in which the early fusion is focused on the fusion of the SEU and the TEU (Figure 7.1), and late fusion combines the body pose and the RGB stream. Moreover, the early fusion considers features, which are extracted from the same feature space (e.g., body pose) whereas in late fusion, the features are combined from separate feature space (Pose and RGB stream). Before the late fusion, features in both the RGB and pose stream are processed through the stream-specific 'Self Multi-Head Attention' followed by a Bi-LSTM. After the late fusion, a GAP and a FC layer is used for the activity classification, as shown in Figure 7.1.

## 7.4 Experiments, Results and Analysis

### 7.4.1 Implementation

For all datasets, a sub-sequence of 20 equally spaced frames is used. For all the pose data, a normalisation step is applied where the data is transformed to body-centred co-ordinates. This is done by subtracting the 'middle of spine' joint from each joint and then normalising with respect to the 'middle of spine' joint. In case of multiple subjects, normalisation is carried out on each subject separately. The video sequences are cropped to a size of 224x224 and the pose sequences are translated accordingly. The model is trained using the Adam optimiser with a fixed initial learning rate of 1e-3 and a decay rate of 1e-6. However, while experimenting with the skeleton model, SGD optimiser has been used with a learning rate of 0.1 as SGD optimiser worked better for pose network. The regularisation factor is set at 1e-5 with $L2$ regularisation. The network has been trained on an Ubuntu PC fitted with an Nvidia Quadro P6000 (24 GB). Mini-batch sizes of 4 were used for 200 epochs to train the model and the categorical cross-entropy is used as a loss function. The proposed model is implemented on Tensorflow with Keras wrapper.

| **Methods** | **Pose** | **RGB** | **Acc** (%) |
|---|---|---|---|
| Part-aware LSTM (Shahroudy et al., 2016) | X | - | 62.9 |
| C3D (Tran et al., 2015) | - | X | 63.5 |
| DSSCA-SSLM (Shahroudy et al., 2017) | X | X | 74.9 |
| Synthesised CNN (Liu et al., 2017) | X | - | 80.0 |
| ST-GCN (Yan et al., 2018) | X | - | 81.5 |
| DPRL+GCNN (Tang et al., 2018) | X | - | 83.5 |
| PDA (Baradel et al., 2018a) | X | X | 84.8 |
| 3Scale ResNet152 (Li et al., 2017a) | X | - | 85.5 |
| Glimpse Clouds (Baradel et al., 2018b) | - | X | 86.6 |
| Proposed Approach (Pose) | X | - | 77.3 |
| Proposed Approach (RGB) | - | X | 85.3 |
| Proposed Approach (Pose+RGB) | X | X | 87.7 |

Table 7.1: Performance of the proposed model and comparison to other state-of-the-art approaches on the NTU RGBD dataset (Shahroudy et al., 2016). All the results are in cross-subject setting which is more challenging than the cross-view setting

### 7.4.2 Experiments and Results

In order to evaluate the performance of the proposed network, three widely used datasets were used: The first two datasets 1) MSR daily Activity (Wang et al., 2012) and 2) NTU-RGBD (Shahroudy et al., 2016) is the same as the datasets used to evaluate the previous model (Chapter 6). In addition, the model presented in this Chapter is evaluated using the 3) SBU Kinect interaction (Yun et al., 2012) dataset. All the datasets also provide 3D pose sequences along with the RGB videos for each action. The standard accuracy metric in percentage is used in all of the evaluations. For NTU-RGBD (Shahroudy et al., 2016) and MSR dataset (Wang et al., 2012) the same standard evaluation protocol as in the last Chapter has been used. The results with NTU-RGBD dataset in Table 7.1 indicate that the proposed model comfortably outperforms other existing models. The proposed model using RGB+pose outperforms the best-performing state-of-the-art model (RGB only in Baradel et al. (2018b)) by 1%. Moreover, the approach (87.7%) is significantly better than the PDA (Baradel et al., 2018a) (84.8%) and DSSCA-SSLM (Shahroudy et al., 2017) (74.9%) approaches that use both pose and RGB information. Furthermore, using RGB only the proposed approach (85.3%) is significantly better than the C3D (Tran et al., 2015) (63.5%) but inferior to the Glimpse Clouds (Baradel et al., 2018b) (86.6%). The Table also shows that only Shahroudy et al. (2017) and Baradel et al. (2018a) have successfully combined RGB and pose information which further shows the importance of the current research.

| **Methods** | **Pose** | **RGB** | **Depth** | **Acc** (%) |
|---|---|---|---|---|
| Ensemble (Wang et al., 2012) | X | - | - | 68.0 |
| Efficient Pose (Eweiwi et al., 2014) | X | - | - | 73.1 |
| Moving Pose (Zanfir et al., 2013) | X | - | - | 73.8 |
| Poselets (Tao; Vidal, 2015) | X | - | - | 74.5 |
| MP (Shahroudy et al., 2017) | X | - | - | 79.4 |
| Actionlet (Wang; Wu, 2013) | X | - | - | 85.8 |
| PDA (Baradel et al., 2018b) | X | X | - | 90.0 |
| Depth Fusion (Zhu et al., 2015) | - | - | X | 88.8 |
| MMMP (Shahroudy et al., 2015) | X | - | X | 91.3 |
| DL-GSGC (Luo et al., 2013) | X | - | X | 95.0 |
| DSSCA-SSLM (Shahroudy et al., 2017) | - | X | X | 97.5 |
| Proposed Approach (Pose) | X | - | - | 76.3 |
| Proposed Approach (RGB) | - | X | - | 90.6 |
| Proposed Approach (Pose+RGB) | X | X | - | 92.5 |

Table 7.2: Comparison of the proposed model with the state-of-the-art approaches on MSR dataset (Wang et al., 2012)

The performance using MSR dataset (Wang et al., 2012) is presented in Table7.2. The proposed model outperforms (92.5%) the PDA approach (Baradel et al., 2018b) (90.0%) using RGB+pose data. Using RGB only, the approach (90.6%) is significantly better than all the approaches that use uni-modal information. The best performing models use a combination of raw depth and pose data which is very memory intensive. Each MSR depth action consumes 45 MB of data while a RGB video requires only around 5 MB. It is not feasible to scale such models to larger dataset. This indicates why many authors have ignored the raw depth-based models in larger datasets like the NTU-RGBD.

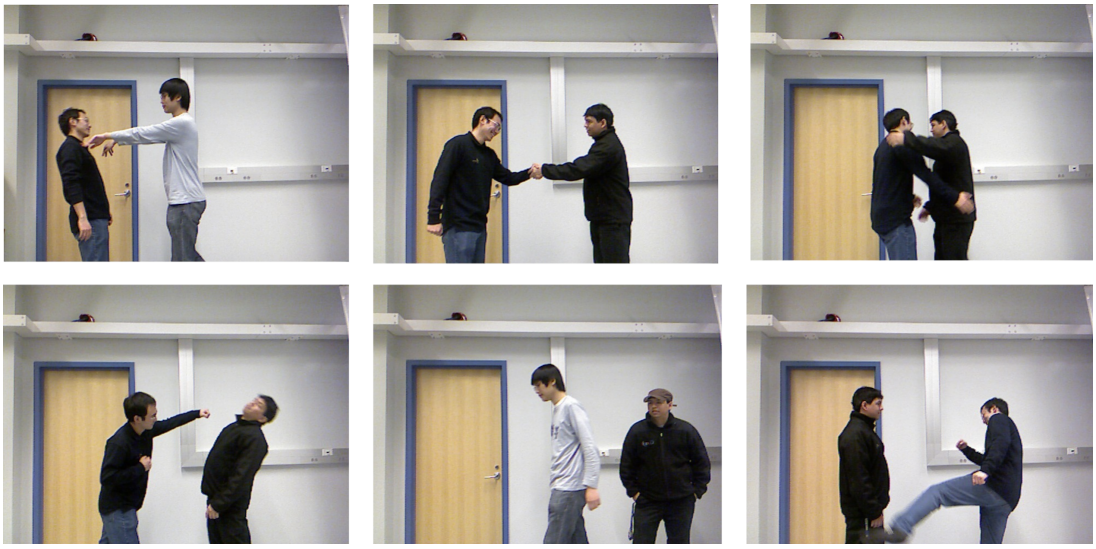


Figure 7.4: Samples from the SBU-Kinect (Yun et al., 2012) interaction dataset. Clockwise from top left: Pushing, Handshake, Hugging, Kicking, Departing and Punching

| **Methods** | **Pose** | **RGB** | **Depth** | **Acc** (%) |
|---|---|---|---|---|
| Joint feature (Yun et al., 2012) | X | - | - | 80.3 |
| Joint feature (Ji et al., 2014) | X | - | - | 86.9 |
| Co-occurence RNN (Zhu et al., 2016) | X | - | - | 90.4 |
| STA-LSTM (Song et al., 2017) | X | - | - | 91.5 |
| ST-LSTM + Trust Gate (Liu et al., 2016b) | X | - | - | 93.3 |
| DSPM (Lin et al., 2016) | - | X | X | 93.4 |
| PDA (Baradel et al., 2018a) | X | X | - | 94.1 |
| VA-LSTM (Zhang et al., 2017b) | - | X | X | 97.5 |
| Proposed Approach (Pose) | X | - | - | 96.2 |
| Proposed Approach(RGB) | - | X | - | 95.5 |
| Proposed Approach (Pose+RGB) | X | X | - | 96.5 |

Table 7.3: Results on the SBU Kinect dataset (Yun et al., 2012). The results shown are the average of five fold cross-validation

The third dataset is SBU Kinect (Yun et al., 2012) interaction dataset, which is a human-human interaction datasets consisting of 282 videos with 8 different activities classes. For evaluation, the authors' protocol of 5-fold cross-validation is followed. The results in Table 7.3 indicate that the current model (96.5%) significantly outperforms the state-of-the-art approaches that use either RGB or pose or their combination. Moreover, the model using pose only (96.2%) is 2.9% better than the best approach that uses ST-LSTM + Trust Gate (Liu et al., 2016b). It is worth mentioning that the performance of the pose model on the SBU Kinect dataset (96.2%) is much better than MSR dataset (76.3 %, Table 7.2). This could be attributed to the train validation split for each dataset. Whereas

in MSR dataset 50% of the data is used for validation, in SBU Kinect only 20% of the data is used for validation. The model sees more data (80%) in case of of SBU Kinect, and hence perform better. Note that this is different from the behaviour of RGB stream which provides better performance in general owing to the pre-trained Inception-ResNet-V2. On the other hand The pose stream is not pre-trained.

### 7.4.3 Ablation study

The ablation study illustrates the impact of the SEU, the TEU and the 'Multi-Head Attention' mechanism in both RGB and pose streams. For RGB stream, a pre-trained Inception-ResNet-V2 (Szegedy et al., 2017) followed by a Bi-LSTM module is used as the base network. This is followed by a GAP and FC layer for training and evaluation. Later the 'Multi-Head Attention' mechanism is included to evaluate its effectiveness. As shown in Table7.4, the 'Attention' mechanism significantly enhances the performance of the RGB stream as compared to the base network. This justifies the significance of the 'Attention' module in the network.

| **Method** | **NTU** | **MSR** | **SBU** |
|---|---|---|---|
| Baseline | 82.2 | 86.9 | 91.7 |
| + Multi-Head Self Attention | 86.6 | 90.6 | 95.5 |

Table 7.4: Experiments show that application of 'Self Multi-Head Attention' mechanism to the RGB network improves the performance significantly. + signifies the addition of that sub-module

In the base network, instead of the SEU and TEU sub-modules, three 1D convolution layers are used. It also does not include the 'Multi-Head Attention' mechanism. For evaluating the pose model, a FC layer is applied on top of the final output of the pose network. Afterwards, SEU is introduced in first three 1D convolutional layers of one of the streams. Then, in the second stream, the TEU is introduced in the first three 1D convolution layers. The 'Self Multi-Head Attention' mechanism is applied to the pose network after the fusion of two streams. Experiments showed that keeping the number of heads at four is optimum for both RGB and pose networks. The Table7.5 shows considerable improvement as a result of SEU, TEU and the 'Multi-Head Attention' mechanism.

| **Method** | **NTU** | **MSR** |
|---|---|---|
| Baseline | 73.3 | 72.5 |
| + SEU | 75.4 | 74.3 |
| + TEU | 75.9 | 75.0 |
| + Multi-Head Self Attention | 77.3 | 76.3 |

Table 7.5: The performance of each network element. + signifies the addition of that sub-module

### 7.4.4 SEU and TEU analysis

This section discusses how the proposed pose encoding is able to exploit the 1D convolutional mechanism to learn and represent more effective and enriched encoding to the network. For processing pose information with 1D convolutions, the 3D pose information for each frame is normally squeezed

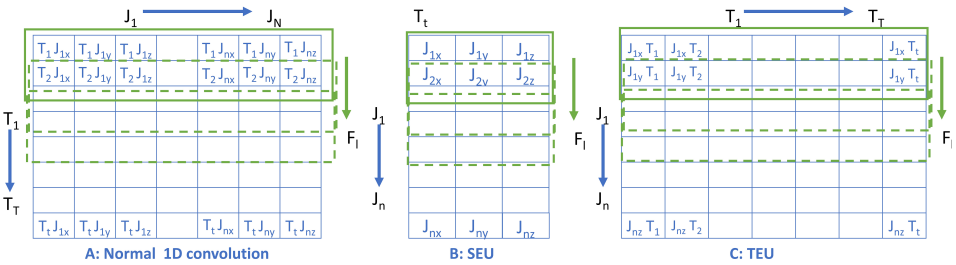



Figure 7.5: Differences in input map for 1D convolutions. a) Normal input maps for 1D convolutions. b) SEU: Per frame input maps of all the joints c) TEU: Whole temporal sequence of each joint is presented as a vector.

into a vector and sequence of frames are stacked to produce a 2D input map as shown in Figure 7.5a. The green rectangles depict a convolutional filter which slides from top to bottom during each stride. This operation has two drawbacks: 1) It does not take into account the spatial relationship among each body joint which is essential for capturing the structural composition and dependencies of the pose attained in each frame. 2) During each convolution stride, only a subset of frames is used. The number of frames used in each stride depends on the filter length. This operation does not effectively capture the long-range temporal dependencies. The proposed SEU and TEU encoding aims to overcome these drawbacks. SEU encodes the pose maps for each frame separately. For each frame, during a convolution operation, a filter (green rectangle in Figure 7.5b) strides over a subset of joints. The operation can be compared to selecting groups of consecutive joints whose interrelations are learned by each filter in a stride. The number of joints in each 'group' is determined by the length of the filter. In addition, the spatial significance of each dimensions is preserved unlike typical input map (Figure 7.5a), the spatial dimension is not squeezed into a vector. The SEU produces maps for each frame which represents the various structural relationships and dependencies of human pose in each frame. On the other hand, the TEU presents the whole temporal sequence for each joint in a vector and stacks all the joints (Figure 7.5c). As a result, the convolution operation learns the whole temporal sequence for a group of joints in a single convolutional stride. Thus, the TEU effectively captures the long-range temporal dependencies for each joint. Figure 7.6 presents the confusion matrix of the classifier for MSR dataset while Figure 7.7 does the same for NTU-RGBD dataset.

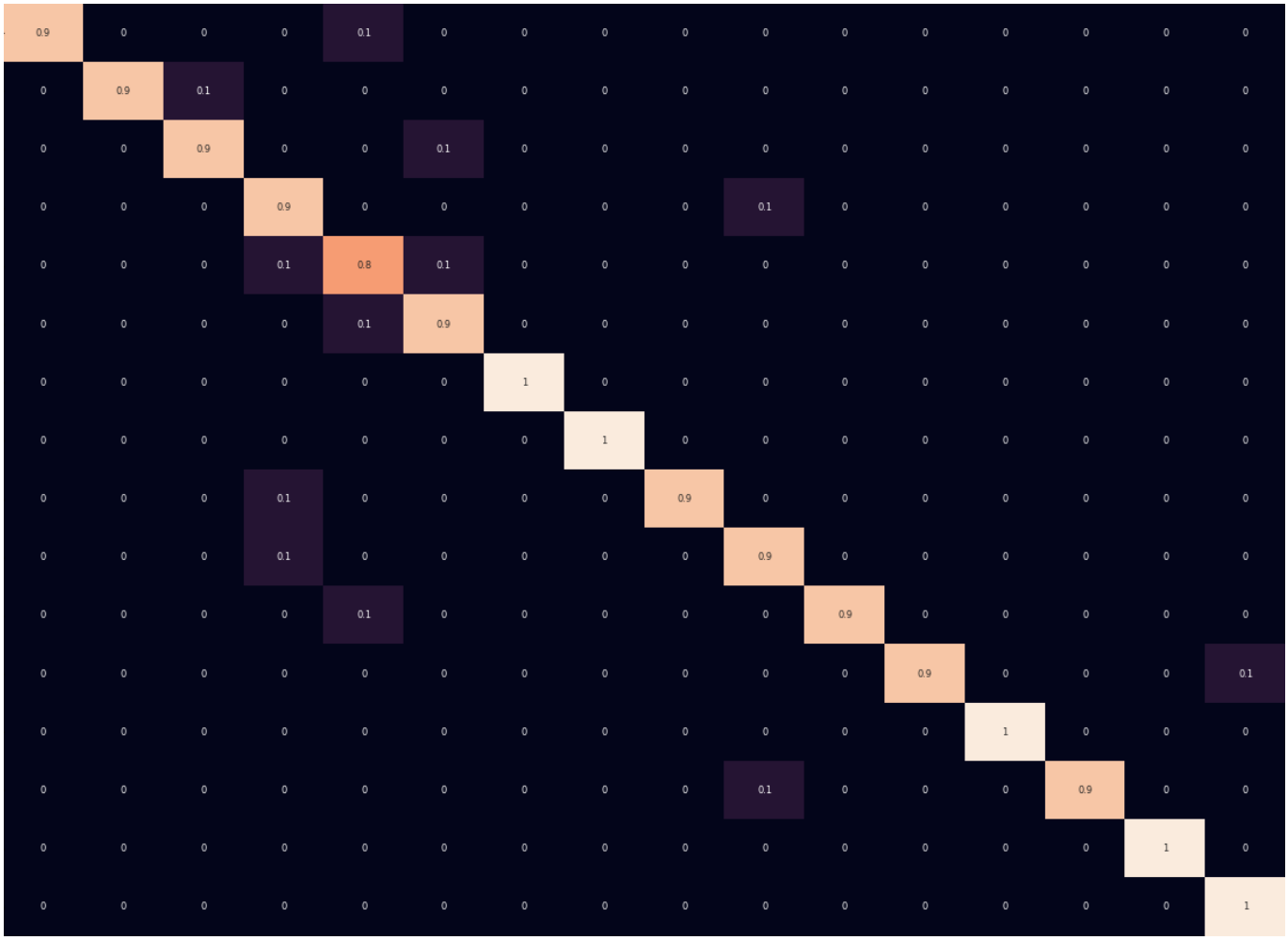



Figure 7.6: Confusion Matrix of the combined monocular video and pose-based classifier with an accuracy of 92.5% (Chapter 6, Table 7.2) for the MSR dataset

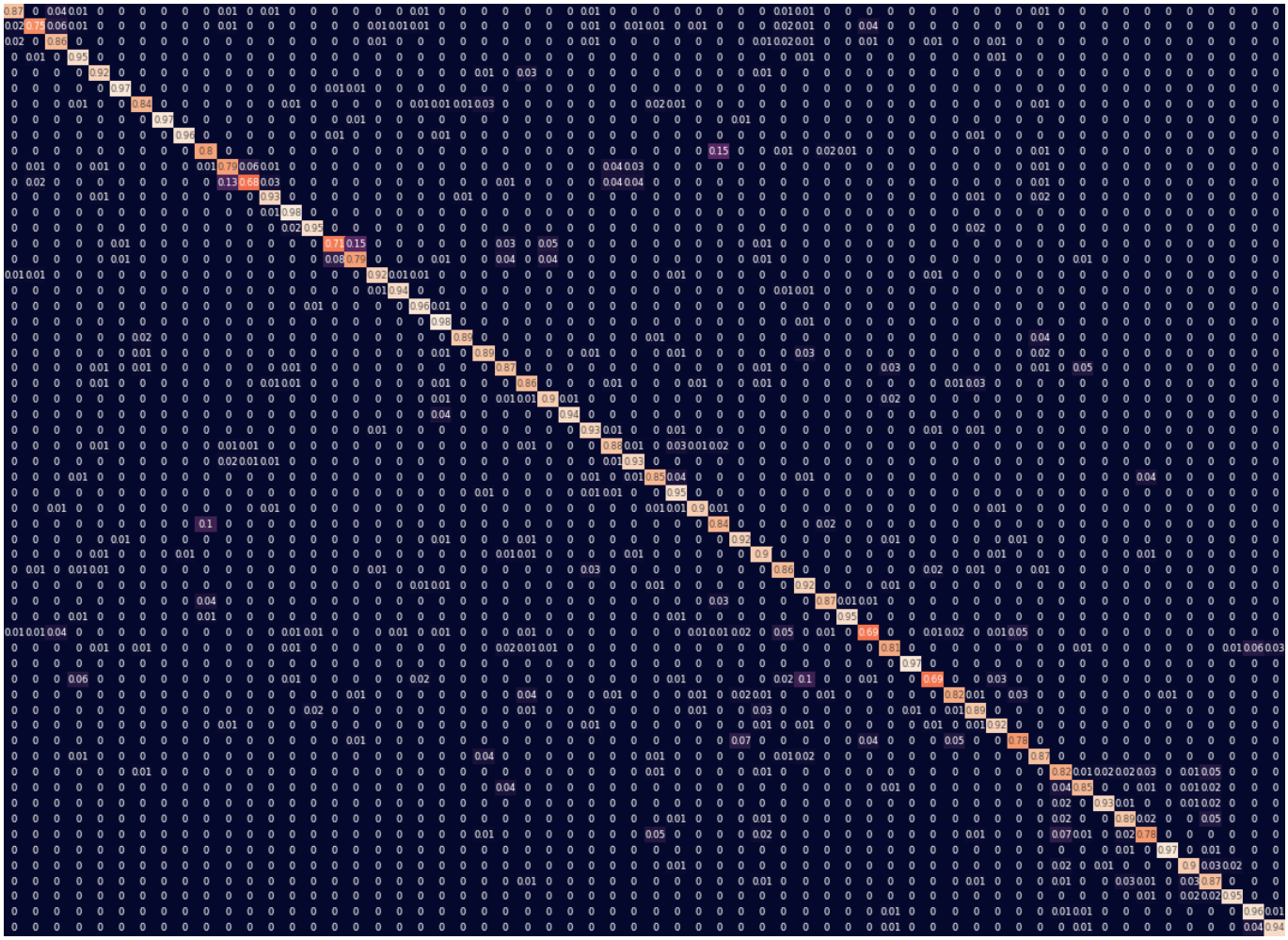


Figure 7.7: Confusion Matrix of the combined monocular video and pose-based classifier with an accuracy of 87.7% (Chapter 6, Table 7.1) for the NTU dataset

## 7.5 Discussion

The model presented in this Chapter is a combined video-pose data model that has achieved better than state-of-the-art results on three datasets (Sec 7.4). The model is significant to the broader literature in two aspects. First, the model introduces enriched feature representation from 3D skeleton sequences that is *'learnt'* instead of hand-crafting such features. This representation is applied to a new multi-stream network that consists of two pose and one video streams. Out of the two pose streams, the first stream presents spatially enriched 3D pose data that captures the structural relationships between the various body joints and learns spatially enhanced representation. The second stream learns the temporal relationship between various time points for each joint individually and presents a temporally enhanced representation. To the best of my knowledge, this is the first model that learns the various structural connections and dependencies from 3D pose information in this manner. As discussed in the literature review (Chapter 3, Sec. 3.4.3), existing models have used hand-crafted features to enhance skeleton representations. Such hand-crafted representations include augmenting coordinates with velocities and acceleration (Demisse et al., 2018; Zanfir et al., 2013), various normalisation techniques for the body joints (Zanfir et al., 2013), and adding relative positions (Ke et al., 2017). Hand-crafting representations require significant time, manual effort and domain knowledge. The benefit of the proposed method is that the model automatically learns

such representations without much effort on part of the developer. The *'learning'* is integrated in an end-to-end trainable manner meaning the model is more generalised and less prone to over-fitting. This learning is carried out through simple 1D convolutions and can be integrated with any pose-based method to enhance the model performance. In the next Chapter, the SEU and TEU are incorporated into the TCN-ResNet (Kim; Reiter, 2017) model based on 1D convolutions. The second impact of this model is that it is able to successfully combine video and pose data for better model performance. Video data offer important context cues, scene information, optical flow information while pose information focuses the network of important spatio-temporal points. Tables 7.1, 7.2 and 7.3 show that only Shahroudy et al. (2017) and Baradel et al. (2018a) have successfully combined RGB and pose information which further shows the importance of this model. From the review of CV-based assessment and rehabilitation methods (Chapter 2), it can be seen that a majority of researchers have used Kinect-based 3D pose information. Also, the literature review (Chapter 3, Sec. 3.4) points towards the increasing availability of pose-data and thus the use of pose-based methods. Thus, integrating the use of pose data becomes more important for ADL. Also, the SEU and the TEU is used to enhance the performance of the multi-label activity recognition model presented in the next chapter. This model has been published in the $25^{th}$ *IEEE ICPR, 2020.*

## 7.6 Conclusion

The activity recognition method presented in this Chapter introduces a novel body-pose encoding method to give us state-of-the-art result for three challenging benchmark datasets. It combines monocular RGB videos and 3D pose estimation whereas the previous Chapter presented a purely RGB video-based method. The next Chapter explores a purely pose-based model for the fifth objective of this study, which is to develop a multi-label activity recognition method. The body-posed encoding method presented in this Chapter is used to improve the performance of the pure pose-based multi-label activity recognition model.

# Chapter 8

# Functional Activity Recognition

## 8.1 Introduction

This Chapter addresses the fifth research objective, which is to design a multi-label ADL recognition model for recognising impairment-specific versions of an ADL. In Chapter 6 (Sec. 6.1), it is discussed that the study approaches the problem of multi-label ADL recognition by focusing on single-label ADL recognition first. Accordingly, the previous two chapters (6 and 7) introduce two different ADL recognition models. These models are evaluated on well-known benchmark datasets which help to demonstrate the model performance against existing state-of-the-art approaches. These benchmark datasets are meant for multi-class single-label classification where there is only one label assigned to a data sample. For example, in multi-class single-label human activity recognition, a video sequence is labelled as either 'Drinking' or 'Walking' and so on. In contrast, in multi-class multi-label recognition, there are multiple labels assigned to each data sample. The dataset presented in Chapter 5 is more suitable for multi-class multi-label classification where there are two labels ('Activity' and 'Impairment') for each data sample. For example, the 'Activity' label could be 'Drinking' whereas the 'Impairment' label could be 'Ataxic'. The model in Chapter 6 uses only monocular RGB images whereas the Chapter 7 presents a combined RGB+pose model. This Chapter in contrast, explores a purely pose-based (human-skeleton) model for multi-label human activity recognition. The activity recognition models of previous chapters (6 and 7) have been utilised to improve the performance of the model presented in this Chapter which is based on TCN-ResNet architecture proposed by Kim; Reiter (2017). This model is then adapted to multi-label classification and trained on the multi-label activity dataset presented in Chapter 5. This enables the model to recognise an ADL and discriminate between the normal and four impairment-specific versions of the same ADL. The main aim of this study is to improve the functional assessment of ADL by discriminating between various impairment-specific versions of the same ADL and this Chapter fulfils this.

The next section highlights the rationale behind the proposed model followed by a brief overview of the TCN-ResNet architecture (Kim; Reiter, 2017). Then, the Chapter presents the proposed approach involving the integration and adaption of SEU, TEU (Chapter 7) and the FV-based pooling mechanism (Chapter 6) to TCN-ResNet. The section further discusses the extension of the model for multi-label activity recognition. This is followed by the experiments, results and analysis section demonstrating the efficacy of the proposed approach. It also includes an ablation study and a qualitative analysis of the model. Finally in the discussion section the contribution of this Chapter with respect to the thesis and the broader literature is discussed.

### 8.1.1 Motivation/Rationale

Researchers have explored monocular RGB video data (Baradel et al., 2018b), pose information (Shi et al., 2019; Baradel et al., 2017), depth information (Zhu et al., 2015) or any combination of these data modalities (Baradel et al., 2018a; Shahroudy et al., 2017; Zhang et al., 2017b) for human activity recognition. Each modality has its own advantages and disadvantages. The introductions in Chapter 6 and 7 explain the rationale for monocular RGB video and combined RGB and pose-based model. However, for processing RGB data, these models use Inception-ResNet-V2 (Szegedy et al., 2017) as the base network which contains around 52 million parameters. In contrast, the TCN-ResNet (Kim; Reiter, 2017) model that forms the basis of the pure pose-based model presented in this Chapter, contains approximately 1.5 million parameters. Bigger networks rely on high-performance GPUs for training and inference which may not be available in clinic or home-based scenarios. The purely pose-based model presented in this Chapter is motivated by the need to explore lightweight models for such scenarios.

The TCN-ResNet architecture originally proposed by Kim; Reiter (2017) has been chosen because of its superior performance compared to the other models on the well-known and challenging NTU-RGBD large-scale dataset (Shahroudy et al., 2016). The model combines TCN (Lea et al., 2017) with residual connections (He et al., 2016) for a pure pose-based activity recognition model. Novelties introduced in the activity recognition models presented in earlier chapters (Chapters 6 and 7) have been utilised to enhance the performance of the TCN-ResNet model. Specifically, the TCN-ResNet model adapts the SEU and TEU components proposed in Chapter 7. In addition, the FV-based learnable pooling mechanism introduced in Chapter 6 is used instead of the GAP layer used towards the end of TCN-ResNet. These modifications lead to significant improvement in the performance of the proposed model compared to the baseline. The model has been evaluated on both the multi-label activity recognition dataset (Chapter 5) and the well-known NTU-RGBD dataset (Shahroudy et al., 2016).

## 8.2 The TCN-ResNet Model

As the name suggests, TCN-ResNet is a combination of TCN (Lea et al., 2017) with residual connections (He et al., 2016). Lea et al. (2017) were the first to introduce TCNs for human activity recognition, which are nothing but a series of 1D convolutions. TCNs provides an alternative to LSTM for processing temporal sequences like human body-pose information (Lea et al., 2017). Unlike LSTM, the processing in TCNs is performed layer-wise where all the time-steps are updated simultaneously (Lea et al., 2017). This results in faster computations in comparison to LSTM where each time-step is processed sequentially (Lea et al., 2017). Lea et al. (2017) uses an encoder-decoder model and do not find skip-connections (residual connections) useful for increasing the model's performance. However, residual connections (He et al., 2016) have been well-studied and used in many well-known 2D CNN-based classification models (Szegedy et al., 2016; Szegedy et al., 2017; He et al., 2016) to improve their performance. In this study, skip-connections have been found to benefit the pose estimation model (Chapter 4), pose-stream (Chapter 7) as well as the base CNN Inception-

ResNet-V2 (Szegedy et al., 2017) for the RGB models (Chapter 6 and 7). In TCN-ResNet, authors are able to exploit residual connections for improving performance of a model based on TCN.

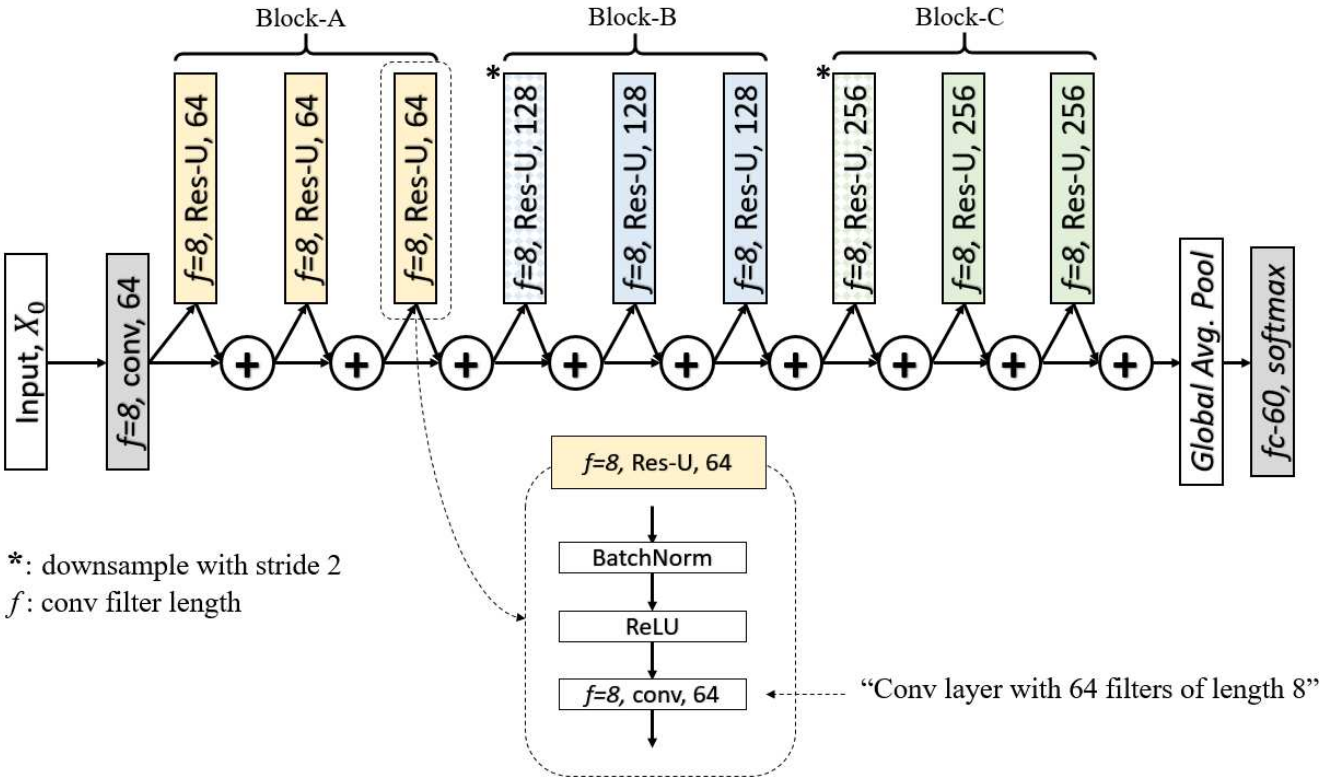



Figure 8.1: The TCN-ResNet (Kim; Reiter, 2017) architecture is a stacking of 1D convolutional layers. There network is divided in three blocks, where layers in each block have same number of filters. The model combines TCN (Lea et al., 2017) with residual connections (He et al., 2016) for a purely pose-based activity recognition model

The architecture of the TCN-ResNet model shown in Figure 8.1, is basically a stacking of 1D convolutional layers followed by the standard GAP + FC layers. As shown, the network is composed of three 1D convolutional blocks (Block-A, Block-B and Block-C) and each of the three blocks consists of three layers of 1D convolutions. Each convolutional operation is followed by a Batch Normalization (BN) and a Rectifier Linear Unit (ReLU) activation function. The convolutional operation at the start of Block-B and Block-C is of stride 2, which means the input is halved along the first dimension (normally time dimension) as it passes from Block-A to Block-B, and then from Block-B to Block-C. In addition to that, the number of filters is doubled from Block-A (64) to Block-B (128) to Block-C (256). There are two paths between any two layers: 1) First is through 1D convolutional operation followed by a BN and a ReLU activation function; 2) Second, through a residual or skip-connection. The two paths are then combined with an addition operation (indicated by '+' sign in Figure 8.1). Let $T$ be the number of frames in a sequence, $J$ the number of body joints, $D$ the number of dimensions of each joint (in this case 3 for 3D pose) and $F$ the total number of filters in a layer. Then, with input vector $V \in \mathbf{R}^{T \times JD}$, 1D convolution operations (Eq. 7.1) in each block performs the following transformation:

$$BlockA : V_{T,J*D} -> M^{a}_{T,F_a} \tag{8.1}$$

$$BlockB : M^{a}_{T,F_a} -> M^{b}_{T/2,F_b} \tag{8.2}$$

$$BlockC : M^{b}_{T/2,F_b} -> M^{c}_{T/4,F_c} \tag{8.3}$$

Here, $F_a = 64, F_b = 128, F_c = 256$ indicate the number of filters and $M^a$,$M^b$, $M^c$ imply the output maps of Block-A, Block-B and Block-C (Figure 8.1), respectively. This architecture of increasing filters and decreasing resolution (in this case temporal resolution) is common to many well-known models (Szegedy et al., 2015; Szegedy et al., 2017; He et al., 2016; Howard et al., 2017), but is different from TCN-based encoder-decoder architecture presented by Lea et al. (2017) in which the resolution is gradually decreased and then increased. Many of the well-known classification architectures have benefited from the residual connections (He et al., 2016; Szegedy et al., 2017) and this (increasing filters and decreasing resolution) could be a reason why residual connections worked for TCN-ResNet (Kim; Reiter, 2017) but did not work for Lea et al. (2017). The output of Block-C (Figure 8.1) is passed through a standard GAP layer followed by a FC layer with Soft-max activation function.

## 8.3 Proposed Approach

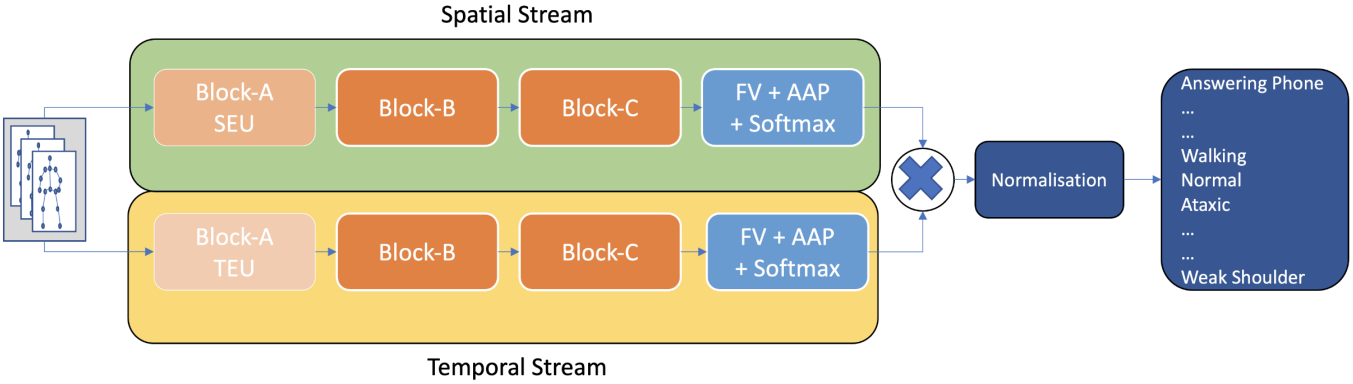



Figure 8.2: The proposed model consists of a spatial and a temporal stream where each stream uses a TCN-ResNet (Kim; Reiter, 2017). Block-A of the spatial stream is used as the SEU (Chapter 7, Sec. 7.3.1) while the same block in temporal stream is used as the TEU (Chapter 7, Sec. 7.3.1). The GAP + FC layer of the TCN-ResNet (Kim; Reiter, 2017) is replaced by a FV-based activity-aware pooling mechanism (Chapter 6, Sec. 6.4.4). The Soft-max output of both the streams are multiplied (indicated by $\times$) and normalised. The model is trained through a multi-hot encoded label wherein each label vector there are two '1's indicating 'Activity' and 'Impairment' labels

In this section first, a brief overview of the model architecture is presented, followed by a detailed description of the individual components. The model presented in this Chapter is an adaptation of the pose-based TCN-ResNet model by Kim; Reiter (2017). The two-stream model consists of a spatial and a temporal stream, as shown in Figure 8.2. The spatial and the temporal stream are each an individual TCN-ResNet (Figure 8.1) excluding the GAP + FC layer. As shown in Figure 8.2, the spatial stream includes the SEU while the temporal stream includes the TEU. The function of the SEU (Chapter 7, Sec. 7.3.1) and the TEU (Chapter 7 Sec. 7.3.2) are the same as in the pose-network of the previous Chapter. In the previous model (Chapter 7), the SEU and TEU are implemented through three 1D convolutional layers. Here, the first block (Block-A) of the TCN-ResNet, which consists of three 1D convolutional layers, is used as the SEU and the TEU in the spatial and temporal streams, respectively. Towards the end of each stream, FV-based activity-aware pooling (Chapter 6, Sec. 6.4.4) is introduced to replace the GAP and the FC layers. Both the streams are then fused together in a very late-fusion manner for an end-to-end trainable multi-label activity recognition

model.

### 8.3.1 Spatial Stream

The rationale behind the spatial stream that includes the TEU is the same as the spatial stream of the pose network presented in the previous Chapter (Chapter 7, Sec 7.3.1). The main design question here was how to adapt TCN-ResNet (Kim; Reiter, 2017) for integrating the SEU. To adapt TCN-ResNet, there were two options: 1) Use part of the TCN-ResNet as the SEU or; 2) Use the whole network as SEU and another one for the rest of the spatial stream. For the first option, the question was whether to allocate Block-A and Block-B or all the three blocks for the SEU. Eqs. 7.2 and 7.3 (Chapter 7, Sec 7.3.1), demonstrate that the size of the output map of the SEU is dependent on the number of filters like normal 1D convolutions (Eq. 7.1). However, the SEU increases the number of filters by a factor of the number of body joints (Eq. 7.2 and 7.3) as compared to normal 1D convolutions. In TCN-ResNet, the outputs of block-B and block-C (Eqs. 8.2 and 8.3) has twice and four times the number of filters respectively, as compared to Block-A. Preliminary experiments were performed with the SEU consisting only of Block-A (Figure 8.1). Including more blocks (Block-B, Block-C) in the SEU increased the number of parameters, making the model slower, while having no positive impact on the performance. The second option of using a full TCN-ResNet for the SEU and another for the rest of the spatial stream also led to increase in number of parameters without any performance benefit. Thus, it was found that utilising the Block-A as SEU and Block-B and Block-C (Figure 8.1) for rest of the spatial stream is the best option. As in the previous section, let $T$ be the number of frames in a sequence, $J$ the number of body joints and $D$ the number of dimensions of each joint (in this case 3 for 3D pose). The input is encoded in a vector $V \in R^{T,J*D}$. The convolutional Block-A (Figure 8.2) is parameterised by the total number of $F_{as}$ filters in each layer. Then, with the help of Eq. 7.2 and 7.3 (Chapter 7, Sec 7.3.1) it is established that the Block-A (Eq. 8.1) when adapted for SEU, performs the following transformation:

$$V_{T,J*D} -> M^{as}_{T,J*F_a} \tag{8.4}$$

Here, $M^{as}$ is the output map of Block-A (Fig. 8.2), which corresponds to $M^a$ in the TCN-ResNet (Eq. 8.1). Thus, instead of a map of dimension $T, F_a$ (Eq. 8.1) the SEU transforms the input to a map of $T, J \times F_a$ (Eq. 8.4). The rest of the spatial stream (Block-B, Block-C Figure 8.2) is the same as in TCN-ResNet with the output of Block-C having size $M^{cs}_{T/4,F_c}$ (Eq. 8.3). Instead of feeding the output of Block-C to the GAP/FC layer as in TCN-ResNet, it is forwarded to the FV-based activity-aware pooling mechanism presented in Chapter 6 (Sec. 6.4.4). The learn-able pooling approach in Chapter 6 exploits the structural information contained withing the hidden Bi-LSTM states. Instead, the learn-able pooling mechanism aims to semantically cluster the output map of 1D-convolutional operation generated by Block-C (Figure 8.2) in this model.

### 8.3.2 Temporal Stream

The temporal stream shown in Figure 8.2 has a functionality similar to the temporal stream of the pose network presented in the previous Chapter (Chapter 7, Sec 7.3.2). A TCN-ResNet (Kim; Reiter, 2017) is used for the temporal stream and the first block (Block-A, Figure 8.1) is adapted for TEU (Figure 8.2). Formally, from Eq. 7.4 and 7.5, (Chapter 7, Sec. 7.3.2) the TEU adapted from Block-A (Eq. 8.1) performs the following transformation:

$$V_{T,J*D} -> M^{at}_{F_a,J} \tag{8.5}$$

Here, $M^{at}$ is the output of Block-A (Figure 8.2) corresponding to Block-A (Eq. 8.1) in TCN-ResNet (Figure 8.1). Thus, the convolution block, Block-B (Figure 8.1), receives an input whose temporal dimension is dependent on the number of filters in Block-A i.e., $F_a$. The output of TEU is passed to Block-B and Block-C (Figure 8.1) in the temporal stream, which performs the following transformation (Eq. 8.2):

$$M^{at}_{F_a,J} -> M^{ct}_{F_a/4,F_c} \tag{8.6}$$

Similar to the spatial stream, the output of the convolution Block-C is fed into the activity-aware learn-able FV pooling mechanism (Figure 8.2) that was presented in Chapter 6 (Sec. 6.4.4).

### 8.3.3 Streams fusion

After spatial and temporal streams, the next design consideration is the mode of fusion of the two streams. The potential points for the fusion of the two streams are at the end of each block. The SEU and the TEU produce maps of different dimension (Eq. 8.4 and Eq. 8.6) at the end of Block-A (Figure 8.1). Moreover, TCN-ResNet (Kim; Reiter, 2017) reduces the temporal dimensions through Block-B and Block-C (Eqs. 8.2 and 8.3). Owing to the above factors the spatial and temporal streams produces maps of different dimensions at the end of each block (Figure 8.2). For example, the output of the spatial stream at the end of Block-C (Figure 8.2) is $M^{cs}_{T/4,F_c}$ (Sec. 8.3.1) and for the temporal stream it is $M^{ct}_{F_a/4,F_c}$ (Eq. 8.6 and Eq. 8.3). This difference does not allow the maps to be fused with either concatenation or addition in a semantic manner. At this stage, the two streams can be fused by flattening and concatenating however, flattening disturbs the spatial and temporal structural organisation of the maps. Moreover, FV-based clustering mechanism relies on such meaningful representations for semantic clustering (Perronnin; Dance, 2007). Empirically, it was observed that flattening the two streams at this stage for fusion lead to poor performance. To preserve the structural information contained in the maps and to cluster them semantically, each stream uses its own FV-based activity aware pooling mechanism. The details of the FV-based pooling mechanism is the same as described in Chapter 6 (Sec. 6.4.4). Thus, the spatial and the temporal streams are fused in a very late-fusion manner where the output of activity-aware pooling is fused together by multiplication followed by normalisation (Figure 8.2).

## 8.4 Training and Evaluation

As with the models presented earlier, this model has been also implemented through Tensorflow framework with Keras wrapper. It was experimentally found that SGD was the best option as optimiser with a learning rate of 0.01. A learning rate scheduler was used where the learning rate was scheduled to drop by 10% on reaching a plateau in training error. A regularisation parameter was set at 1e-5 with $L2$ regularisation. The model was implemented on a 24 GB Nvidia Quadro P6000 GPU with a batch size of 16. To train the model for multi-label classification 'binary cross-entropy' was used as loss function instead of the 'categorical cross-entropy' which is normally used for single-label classification. Further, a custom error metric was implemented where ground-truth was presented as multi-hot encoded labels. Two separate one-hot encoded labels prepared as 'Activity' labels and 'Impairment' labels were concatenated to form the final ground truth labels. Let there be $A$ activity classes and $I$ impairment classes. For $a_{th}$ activity class where $a \in \{1...A\}$ and $i_{th}$ impairment class where $i \in \{1...I\}$, the one hot-encoded labels for activity and impairment respectively are:

$$AL_{m\in A} = \begin{cases} 1, & \text{if } m = a, \\ 0, & \text{if } m \neq a. \end{cases} IL_{n\in I} = \begin{cases} 1, & \text{if } n = i, \\ 0, & \text{if } n \neq i. \end{cases} \tag{8.7}$$

To create the final ground truth label $GT$, the two labels were simply concatenated:

$$GT = AL_{m\in A} \oplus IL_{n\in I} \tag{8.8}$$

Thus, each of the ground truth label vectors $GT$ had two '1' values indicating activity and impairment. In $GT$, the 'Activity' label came from the first $A$ elements whereas the 'Impairment' label was determined from final $I$ elements. Thus, to evaluate the model, the prediction probability vector (i.e., the model output) was split into two parts where the first part contained the first $A$ elements indicating the 'Activity' class probabilities and the rest $I$ elements indicated the 'Impairment' probabilities. Afterwards, the accuracy for the 'Activity' and 'Impairment' was calculated individually in a normal manner. Finally, prediction by the model was considered to be true if both the 'Activity' and 'Impairment' predictions were correct.

## 8.5 Experiments Results and Analysis

| Model | Mode | E2E | RI | CS(%) |
|---|---|---|---|---|
| TCN-ResNet (Kim; Reiter, 2017) | P | ✓ | ✓ | 74.3 |
| Synth. CNN (Liu et al., 2017) | P | **x** | **x** | 80.0 |
| ST-GCN (Yan et al., 2018) | P | ✓ | ✓ | **81.5** |
| DPRL+GCNN (Tang et al., 2018) | P | **x** | **x** | 83.5 |
| PDA (Baradel et al., 2018a) | RP | ✓ | **x** | 84.8 |
| 3scale-ResNet152 (Li et al., 2017a) | P | ✓ | **x** | 85.0 |
| Glimpse Clouds (Baradel et al., 2018b) | R | ✓ | **x** | 86.6 |
| Fisher Vectors (Chapter 6) | R | ✓ | **x** | 87.2 |
| Learned-Encoding (Chapter 7) | RP | ✓ | **x** | 87.7 |
| DGNN (Shi et al., 2019) | P | **x** | ✓ | 89.9 |
| Proposed Model | P | ✓ | ✓ | 80.2 |

Table 8.1: The proposed model achieves competitive accuracy when compared with other pose-based state-of-the-art models given the constraints of data mode (P: Pose, R: RGB-video), being end-to-end trainable (E2E) and random initialisation (RI). Given these constraints ST-GCN achieves the best performance and the model achieves performance close to ST-GCN.

This section describes the experiments that were performed to evaluate the efficacy of the proposed model. The proposed pose-based model is evaluated in both single-label and multi-label mode. Evaluation in single-label mode with publicly available benchmark dataset allows the model to be compared with existing state-of-the-art models. Similar to the earlier chapters (Chapter 6 and Chapter 7), the well-known and challenging NTU-RGBD (Shahroudy et al., 2016), which contains around 60K samples distributed over 60 action classes, has been used for evaluation in single-label mode. Also, similar to the previous models, the authors' protocol of cross-subject (CS) evaluation has been used. This protocol uses different subjects for training and evaluation and is harder than the cross-view protocol, which uses different views but the same subjects for training and validation. Table 8.1 compares the proposed model to existing state-of-the-art approaches and shows that the model achieves competitive performance under the constraints of data modality (pose-based, RGB video-based), end-to-end trainability and random initialisation (i.e., not pre-trained). Clearly, the proposed model comprehensively outperforms the TCN-ResNet (Kim; Reiter, 2017) baseline. The current model has the advantage of being end-to-end trainable as compared to Shi et al. (2019), Tang et al. (2018), and Liu et al. (2017). Also, in contrast to Baradel et al. (2018b), Li et al. (2017a), Liu et al. (2017), and Tang et al. (2018), the current model does not pre-train the proposed model which reflects the true capacity of a model to learn without prior information. Given these constraints the best performance is achieved by ST-GCN (Yan et al., 2018) and the proposed model achieves

almost similar performance while being a very lightweight model. ST-GCN (Yan et al., 2018) uses 8 Nvidia Titan X GPUs for training while the current model requires the equivalent of only one Titan X GPU. CV-based assessment of physically impaired persons are often home or clinic-based where high-performance GPUs may not be feasible and therefore, it is necessary for a model to be lightweight.

| **Model** | **Mode** | **A** | **I** | **Final** |
|---|---|---|---|---|
| I3D (Carreira; Zisserman, 2017) | R | 87.2 | 65.9 | 55.9 |
| C3D (Tran et al., 2015) | R | 90.1 | 73.2 | 63.3 |
| TCN-ResNet (Kim; Reiter, 2017) | P | 91.2 | 69.0 | 63.4 |
| Propose Model (single-label) | P | - | - | 76.7 |
| Proposed Model (multi-label) | P | 97.1 | 80.7 | **78.8** |

Table 8.2: Evaluation of the proposed dataset using different methods. For each sample, the models predict 'Activity' (A) and 'Impairment' (I) and a model's prediction is considered correct if both the 'Activity and 'Impairment' predictions are true. In single label mode each 'Activity-Impairment' combination was allocated a unique label. Mode: Pose (P), RGB-video (R)

Table 8.2 shows the performance of the proposed model on the multi-label activity recognition dataset (Chapter 5) and compares it with the performance of TCN-ResNet (Kim; Reiter, 2017). The last section (Sec. 8.4) describes that the accuracy for 'Activity' and 'Impairment' is calculated separately and combined to get the final accuracy. Thus, the Table 8.2 shows the model performance for 'Activity' and 'Impairment' labels individually and then shows the 'Combined' accuracy. As discussed, in the last section (Sec. 8.4), the combined accuracy considers the model prediction to be correct when both the 'Activity' and 'Impairment' labels are predicted correctly. The following protocol is used to evaluate the current model on the multi-label activity recognition dataset. A cross-validation approach is used where the dataset is split into two subject-wise folds for good generalisation. Cross-validation tests a model's ability to generalise and minimise the effect of sample-bias (Bishop, 2006). The first fold uses subjects 1, 3, 5, 7, 9 for training while subjects 2, 4, 6, 8, 10 are used for validation and vice-versa for the second fold. Thus, out of 5685 samples, the dataset is split into two groups of approximately equal groups of 2869 and 2816 samples which indicates a very good generalisation protocol. This subject-wise split of using half of the subjects for training is inspired by the well-known MSR daily activity dataset (Wang et al., 2012). Customarily for machine learning models, five-fold cross-validation is used (Bishop, 2006), but here the dataset was split into two parts following the well-known MSR dataset (Wang et al., 2012). Splitting the dataset into two parts indicates better generalisation as opposed to five-fold cross-validation where 80% of the dataset is used for training and rest 20% is used for validation. In order to understand the model's true capacity to learn, the model was randomly initialised and no transfer learning was used. The results in Table 8.2 are the weighted averages of the two-fold cross-validation mentioned above. Experiments were also performed in single-label mode where each 'Activity-Impairment' combination was allocated a unique label making a total of 50 classes. The results in Table 8.2 show that the current model comprehensibly out-performs, the base TCN-ResNet model, in both single-label and multi-label mode.

It is also important to note that the multi-label classification performs better than the single label classification. The table also shows that the model performs better than two well-known RGB video-based models, the I3D (Carreira; Zisserman, 2017) and the C3D (Tran et al., 2015). Also, please refer to Appendix A.2 for the confusion matrix of the model.

### 8.5.1 Ablation study

| **Model** | **Split 1** | | | **Split 2** | | | **Weighted-Average** | | |
|---|---|---|---|---|---|---|---|---|---|
| | **A** | **I** | **Final** | **A** | **I** | **Final** | **A** | **I** | **Final** |
| TCN-ResNet (Kim; Reiter, 2017) | 90.4 | 72.0 | 65.3 | 90.0 | 69.0 | 61.9 | 90.2 | 70.5 | 63.6 |
| Two-Stream (TEU) | 92.9 | 73.1 | 69.3 | 96.2 | 75.7 | 73.3 | 94.6 | 74.4 | 71.3 |
| Two-Stream (TEU + SEU) | 93.1 | 74.4 | 69.8 | 95.0 | 79.1 | 75.6 | 94.0 | 76.8 | 72.7 |
| Two-Stream (SEU + TEU) + FV | **96.9** | **77.7** | **75.3** | **97.3** | **83.7** | **82.2** | **97.1** | **80.7** | **78.8** |

Table 8.3: Ablation study demonstrating the effectiveness of the two-stream architecture and FV-based activity-aware pooling in multi-label model

In this section, an ablation study is presented that demonstrates the impact and effectiveness of various components of the proposed model. Table 8.3 demonstrates the evolution of the model from the base TCN-ResNet to the final model through step-by-step inclusion of spatial-temporal architecture and the FV-based pooling. First, experiment is conducted with the original TCN-ResNet (Row 1), which gives a final accuracy of 63.6%. Then the two-stream architecture consisting of two parallel TCN-ResNets (Row 2) is introduced. Here, TEU is introduced to Block-A (Figure 8.2) of one of the streams making the stream temporal in nature. This greatly improves the accuracy which is further enhanced by the introduction of SEU to the spatial stream (Row 3). The final model accuracy is greatly enhanced by the introduction of FV-based pooling to both the streams (Row 4).

| **Model** | **NTU** | **Split 1** | **Split 2** | **Weighted-Average** |
|---|---|---|---|---|
| TCN-ResNet (Kim; Reiter, 2017) | 74.3 | 61.2 | 66.5 | 63.8 |
| Two-Stream (SEU+TEU) | 77.4 | 63.2 | 69.5 | 66.3 |
| Two-Stream (SEU + TEU) + FV | **80.2** | **73.4** | **80.0** | **76.7** |

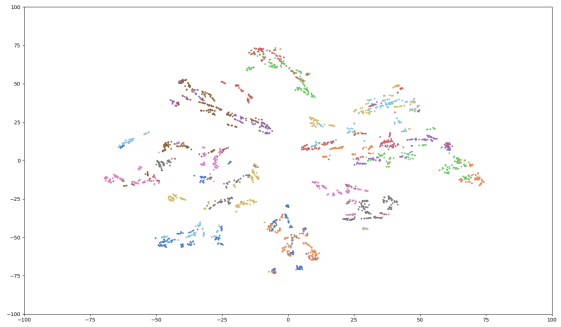
Table 8.4: Ablation study demonstrating the effectiveness of the two-stream architecture and FVs in single label mode

Ablation studies were also carried out in single-label mode to further prove the impact of the two-stream architecture and FV-based learn-able pooling mechanism on the current model. The second column shows the results on the NTU-RGBD dataset (Shahroudy et al., 2016) while the other three columns indicate the results on the current dataset (Chapter 5). As in Table 8.2, for single-label

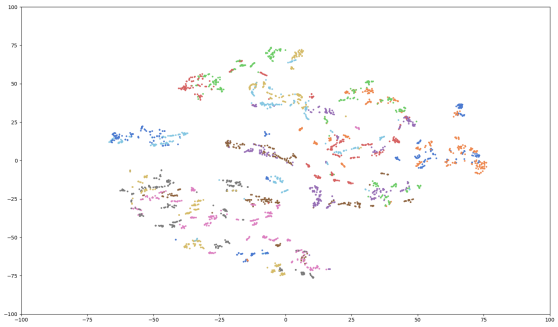
experiment each of the 50 'Activity-Impairment' combination was allocated a unique label. The performance of base TCN-ResNet (Table 8.4, Row 1) is similar to the performance of multi-label case (Table 8.3, Row 1). The results (Table 8.4, Row 1) also show that the two-stream architecture including the SEU and TEU improves the performance of the model although the impact of the same is greater in case of multi-label classification (Table 8.3, Row 3). However, there is a massive improvement in performance when FV-based activity-aware pooling mechanism is incorporated into the two-stream architecture (Table 8.4, Row 3). Whilst the impact of FV on the overall model performance is only marginal in the video-based model in Chapter 6 (0.6%) and NetFV (Miech et al., 2017) (1%), it is much more significant in the present pose-based model. In the present model, FV improves the performance by 6% with multi-label and by 10% with single label supervision. This could be due to the two-stream spatial-temporal architecture of the network. This architecture allows the FV to work on a richer spatial and temporal representation of the input data as illustrated in the next section. Preliminary experiments were also carried out with FV on the base TCN-ResNet (Kim; Reiter, 2017) where there was no impact on the model performance. This further indicates that FV performs better with the two-stream spatial-temporal architecture. The next section further elaborates the impact of FV based activity-aware pooling on the two-stream architecture through t-SNE visualisation and Davies Bouldin Index (DBI).

### 8.5.2 Analysis

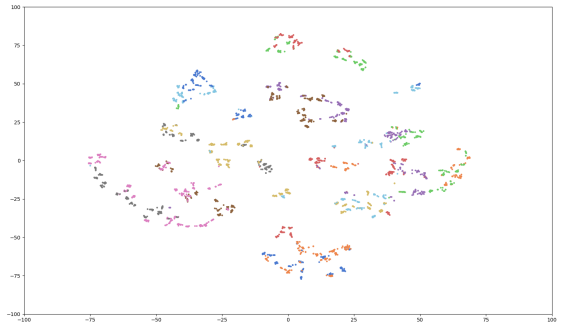
In addition to the ablation study, the model was also studied with t-SNE (Maaten; Hinton, 2008) algorithm with DBI (Davies; Bouldin, 1979) to further analyse its discriminating capability. While t-SNE algorithm is basically a dimensionality reduction technique, it is more often used to visualise high dimensional data into a 2D or 3D map to qualitatively illustrate a model's efficacy (Maaten; Hinton, 2008). On the other hand, DBI, which is a measure of cluster separability, quantitatively indicates how good the cluster separation is. It is the ratio of intra-cluster distance to inter-cluster distance and a lower value indicates better cluster separation. Figure 8.3 shows three models corresponding to the cases presented in Table 8.4. Figure 8.3a shows the base TCN-ResNet case as in row 1 (Table 8.4) and Figure 8.3b shows the two-stream model as in row 2 of Table 8.4. As in Table 8.4, in both the cases FV is not used. Instead, the output is taken from the layer before the GAP layer present towards the end of TCN-ResNet (Figure. 8.1). In the case of Figure 8.3c, which corresponds to row 3 of Table 8.4, the points are from the output of FV (i.e., before the final activity-aware pooling layer) (Figure 8.2 and 6.2). In the current model, the spatial and temporal streams are combined in a late fusion manner after the Soft-max layer. Thus, each stream has its own FV output (Figure. 8.2) which is extracted separately and concatenated for final visualisation (Figure 8.3b and 8.3c). The visualisation clearly shows a better clustering in the case of FV (Figure 8.3c) than without it (Figure 8.3a and 8.3b). This is also backed by much better DBI which is less than half (2.60) for FV than without it (5.65 and 5.55) as indicated in Table 8.5c.



(a) TCN-ResNet (Kim; Reiter, 2017)



(b) Two-stream model (SEU + TEU)



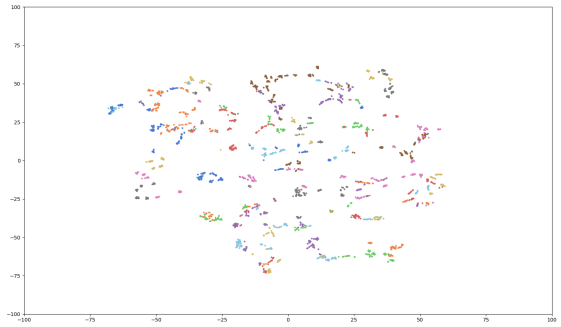
(c) Two stream model (SEU + TEU) + FV

| Figure | Model | Score |
|---|---|---|
| Fig. a | TCN-ResNet | 5.65 |
| Fig. b | Two stream (SEU + TEU) | 5.55 |
| Fig. c | Two stream (SEU + TEU) + FV | 2.49 |

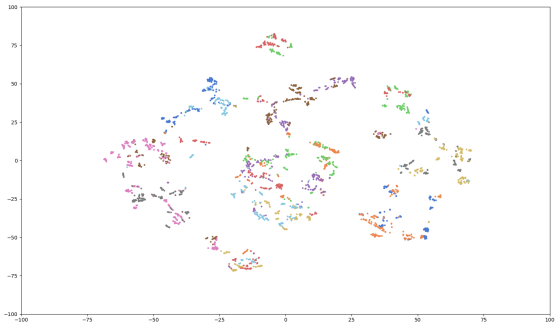
(d) Davies-Bouldin Index

Figure 8.3: t-SNE plot of the output of the layers before the final pooling. a) Output of base TCN-ResNet before GAP. b) No FV: Output of the two-stream model taken before GAP layer in each stream and concatenated. c) FV: Output of the two-stream model taken from FV before activity aware pooling layer in each stream. d) The corresponding DBI score. A lower score indicates better clustering. X-axis: Dimension 1, Y-axis: Dimension 2

In the visualisation, it is not very apparent that the two-stream model (Figure 8.3b) is better than the original TCN-ResNet (Kim; Reiter, 2017) although DBI (Table 8.5c) of 5.55 does indicate that the two-stream architecture is marginally better than TCN-ResNet's 5.65. This could be because for visualisation purpose, the spatial and temporal stream outputs are simply concatenated whereas in the actual model the two streams are element-wise multiplied and then normalised. The output of the two streams before their respective GAP layer have different dimensions. Therefore, element-wise multiplication for visualisation purpose is not possible.



(a) Temporal stream without FV



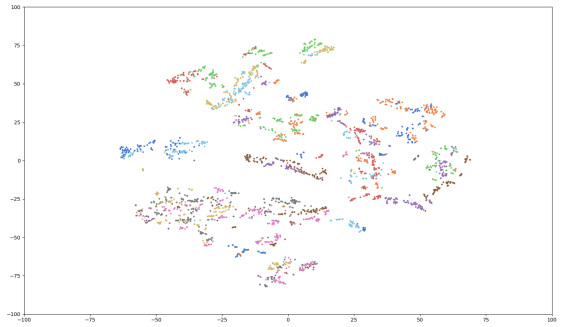
(b) Temporal stream with FV

| Figure | Model | Score |
|---|---|---|
| Fig. a | Temporal stream no FV | 39.98 |
| Fig. b | Temporal stream with FV | 12.22 |

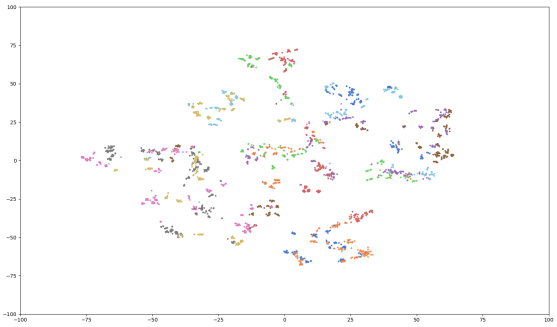
(c) Davies-Bouldin Index

Figure 8.4: Comparison of temporal streams without and including FV through t-SNE (a, b) and corresponding DBI (c)

Furthermore, it is a fact that the temporal stream by itself does not provide very good discriminability but rather it complements the spatial stream. During preliminary experiments it was found that the temporal stream on its own provides very poor performance and hence was excluded from ablation studies in standalone mode. This is also evident from both cluster visualisation and the DBI shown in Figure 8.4 and Table 8.4c. The DBI of the temporal stream with or without FV (Table 8.4c) is much lower than the original TCN-ResNet (Kim; Reiter, 2017) (Table 8.3d, Row 1). However, the same t-SNE visualisation (Figure. 8.4) and the corresponding DBI (Table 8.4c) proves that the performance of the temporal stream is comprehensively improved when FV is introduced to the stream.



(a) Spatial stream without FV



(b) Spatial stream with FV

| Figure | Model | Score |
|---|---|---|
| Fig. a | Spatial stream no FV | 6.28 |
| Fig. b | Spatial stream with FV | 2.60 |

(c) Davies-Bouldin Index

Figure 8.5: Comparison of spatial streams without and including FV through t-SNE (a, b) and corresponding DBI (c)

Similarly, Figure 8.5 including Table 8.5c show that the FV also significantly improves the performance of the spatial stream. It is interesting to note that the DBI score for temporal streams both with or without FV is lower than the base TCN-ResNet. But, when combined with their corresponding spatial streams the resultant model provides better performance than the original TCN-ResNet (Figure 8.3). Continuing with the analysis, the following discussion illustrates the impact of cluster sizes on the model performance.

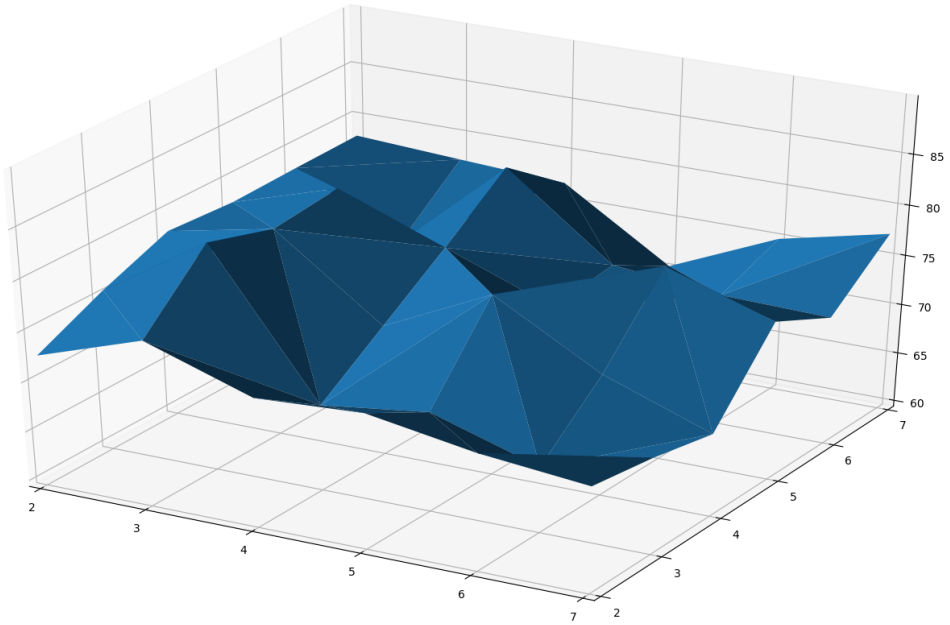



Figure 8.6: Grid search for appropriate cluster-sizes show several parameter choices provides close to peak performance. This indicates that the TCN maps can be semantically clustered in multiple ways. Search range: $2^n$, where $n = 2, 3, 4, 5, 6, 7$

The number of clusters in FV is a tune-able hyper-parameter. A grid-search was performed within the search space $2^n$, where $n = 2, 3, 4, 5, 6, 7$ to understand the impact of cluster sizes on the model performance. The search results illustrated in Figure 8.6 which show that there are several peaks indicating higher performance with multiple cluster-size settings. The best performance (78.8%) is obtained with a cluster-size (CS) of $2^3$ for both the streams. Similar, results (78.0%) are obtained with CS is set at 8 (Spatial) and 16 (Temporal). CS of 16 (Spatial) and 64 (Temporal) gave 78.6% while CS of 64 (Spatial) and 32 (Temporal) gave 77.2% accuracy. The results suggest that the TCN maps can be semantically clustered in a more than one way. In this model, the best performing cluster size of $2^3$, for both spatial and temporal stream is on the lower side of the tested range from $2^2$ till $2^7$. Table A.2 (Appendix A.2) shows the outcome of the grid search in further details.

### 8.5.3 Confusion Matrices


Figure 8.7: Activity Confusion Matrix


Figure 8.8: Impairment Confusion Matrix

In a purely pose-based model, the role of Kinect-based pose measurement becomes vital. The model has no other data such as scene, contextual information apart from the human body-pose extracted from Kinect. As discussed in Chapter 2, Kinect has been very widely used for pose-based models. Thus, researchers have also focused on reliability and accuracy of pose information produced by Kinect. Otte et al. (2016) found that accuracy of Kinect was moderate to optimum depending on dimension, landmark location and performed task. Yang et al. (2015a) found that Kinect accuracy was good if the object was positioned within certain range. In this study, experimentally it was found that 1 to 5 meters was the optimum range for accurate pose estimation under the current setup. This was maintained throughout the dataset collection. Galna et al. (2014) showed that the Kinect was accurate for gross spatial characteristics of clinically relevant movements. However, the spatial accuracy of smaller movements was not the same. This could be one of the reasons for the relatively less accuracy of pose-based model in the current study. The confusion matrix for the 'Activity' classification (Figure 8.7) and results Table 8.2 show that the accuracy is very good (97.1%). This could be partly attributed to the distinctness in activity classes which the model is able to properly capture. However, it can be also attributed to the pose estimation accuracy of Kinect which is sufficiently accurate for grossly distinct spatial movements. In contrast the the 'Impairment' confusion matrix (Figure 8.8) and Table 8.2 show that the accuracy of 'Impairment' classification is much less (80%) which is responsible for the overall low performance of the model. This could be partly due relatively less pose estimation accuracy for similar but subtly different spatial movement for various impairments within the same ADL. For example, 'Clapping-Normal' is very similar to 'Clapping-Tremors' except for the tremor part in which the hands shake as they move. The difference in spatial movement in this case is much more subtle in contrast to distinct activity classes. However, Table 8.2 shows that the pose models has performed much better than

some of the existing state-of-the-art video-based model which justifies the use of Kinect-based pose information for functional ADL recognition.

The 'Activity' Confusion matrix in Figure 8.7 shows that the model has performed very well for all the classes except for class 5 which is 'Drinking' (Table 5.1). The model has confused the 'Drinking' ADL with class 10 'Wearing Glasses' (Table 5.1). These two ADL are very similar with the hand moving towards the face in both the cases. In contrast, the 'Impairment' confusion matrix is relatively more random. The classes 'Normal', 'Shoulder Weakness' and 'Wider Gait' have relatively better performance (over 90%), while the other four impairments stand well below 80%. 'Tremors' has the worst performance (64%) which the model confuses with 'Ataxic' in 23% of the cases. 'Ataxic' and 'Tremors' are somewhat similar in the sense that the hand shakes in both the cases. This suggests that the model focuses more on hand poses. Reciprocally, the model confuses 'Ataxic' with 'Tremors' in 17% of the cases. 'Elbow rigidity' is the other poor performing class which the model confuses with 'Normal' (8%) , 'Ataxic' (10%) and 'Tremors' (8%) which are the other upper body actions. Similarly, 'Knee Rigidity' is a lower body action which the model confuses with other impairments involved with the lower body. Thus, the model is unable to predict 'Elbow Rigidity' and 'Knee Rigidity' with other upper and lower body impairments with no particular bias towards any particular impairment.

### 8.5.4 Complexity Analysis

| **Model** | **Trainable Parameters** ($10^6$) | **FLOps** ($10^9$) | **Inference time (ms)** |
|---|---|---|---|
| TCN-ResNet (Kim; Reiter, 2017) | 1.838 | 0.769 | 1.18 |
| Two-Stream (SEU+TEU) | 4.596 | 5.412 | 1.63 |
| Two-Stream (SEU + TEU + FV) | 4.992 | 5.414 | 1.68 |

Table 8.5: Complexity analysis of the model in terms of millions of parameters ($10^6$), billions ($10^9$) of FLOps and inference time in milliseconds (ms).

Model complexity is represented as the number of trainable parameters, duration of forward pass or inference time and Floating Point Operations (FLOps). The number of trainable parameters 1.84 million (Table 8.5) in the base or original TCN-ResNet (Kim; Reiter, 2017) makes it a very lightweight network. Table 8.5 shows that model complexity in terms of FLOps and number of parameters increase when the model SEU and TEU is introduced to the model in a two-stream architecture. With the introduction of the two-stream architecture the number of trainable parameters increases by almost 2.5 times whereas the number of FLOps increases by seven times. This results in an increase of inference time for the model by around 1.4 times. As shown in the final row of Table 8.5, the introduction of FV only marginally increases the number of trainable parameters. As a result, there is a slight increase in inference time by 0.1 milliseconds. The introduction of two-stream architecture, including the SEU and the TEU increases the multi-label accuracy by around 9% (Table 8.3) while for single-label accuracy the increase is around 3.5% (Table 8.4). This causes the model to have 2.5 times more parameters. Whereas, FV increases the model complexity including inference times

marginally (Table 8.5), it is able to increase the performance by 6.1% (Table 8.3) and 9.7% (Table 8.4) for multi-label and single label classification respectively.

## 8.6 Discussion

This Chapter presents a multi-label activity recognition model which is the fifth and final objective of this study. As mentioned earlier (Sec. 8.1), in multi-label activity recognition there is more than one label that needs to be predicted for each data sample. When this model is trained on multi-label human activity dataset (Chapter 5), the model is able to discriminate between various normal and impaired ways of performing the same ADL. In the domain of CV-based physical assessment, it is important to recognise normal ADL from impaired versions (Chapter 1, Sec. 1.1.2), but the CV community has not fully explored the same (Chapter 2, Sec. 2.11). Together with the multi-label dataset (Chapter 5), the model presented in this Chapter is a stepping stone in this direction. On the other hand, in the area of CV focusing on DL-based activity recognition, multi-label activity recognition is yet to be fully explored (Chapter 3, Sec 3.4.2). To the best of my knowledge, this study is the first to contribute towards multi-label human activity recognition. In the current model, multi-label activity recognition performs slightly better than single-label recognition, where each 'Activity-Impairment' combination was given a unique label (Table 8.2). Whereas through this dataset (Chapter 5) only two-label ('Activity' and 'Impairment') classification was explored, in practice there could be more than two labels. For example, it is not uncommon for physically impaired persons to have more than one condition. It is reasonable to assume that with more labels per sample, multi-label activity recognition will become more significant while performance through single-label supervision may suffer.

In this work, three different data modalities for human activity recognition has been explored. This Chapter presents a pose-based model while the previous two Chapters introduced pure RGB (Chapter 6), and combined RGB+pose-based (Chapter 7) models. Each modality has its own advantages and disadvantages. RGB videos contain a lot of information regarding the scene, as well as objects handled by subjects can provide contextual information, which is vital in discriminating various human activities. On the other hand, 3D body-pose information contains 3D location of human body joints and/or body parts for each frame. This, sequential information helps a combined RGB+pose-based model to focus the network on more important spatio-temporal structures in videos (Baradel et al., 2018a). However, RGB videos consume a lot of space and are very memory intensive and the same is true for depth data. The RGB videos from the NTU-RGBD (Shahroudy et al., 2016) dataset consume around 136 GB, while the RGB videos from the multi-label dataset (Chapter 5) occupy 57 GB. Similarly, depth videos from the NTU dataset consume 886 GB while the raw depth videos from the dataset presented in this study take up more than 1 Terabytes of space. In comparison to that pose information from NTU dataset (Shahroudy et al., 2016) and the multi-label activity dataset (Chapter 5) consume only around 13 and 3 GB of space respectively. In addition to storage requirements, RGB-based models are also expensive in terms of memory requirements. The RGB-based model presented in Chapter 6 contains around 54 million parameters, while the RGB stream in Chapter 7 contains around 52 million parameters. In comparison to that the pose-stream in Chapter

7 contains less than a million parameters while the model in this Chapter has 5 million parameters only. Thus, recently researchers have focused on pose-based models and some of the best performing models for human activity recognition are pose-based (Chapter 3, Sec 3.4.3). The results in Table 8.1 show that the performance of the proposed pose-based model on the NTU-RGBD (Shahroudy et al., 2016) is much less than the current state-of-the-art. However, as highlighted in the Table, given the constraints of random initialisation and end-to-end trainability, the current model is quite close to the best performing model by Yan et al. (2018). The ability to be end-to-end trainable makes a model convenient for tasks such as real-time use. The constraint of random initialisation reflects a model's true capacity to learn without prior knowledge. While the current pose-based model is based on TCN, there has been a significant improvement in the performance of posed-based model lead by graph-based neural networks (Yan et al., 2018; Tang et al., 2018; Shi et al., 2019). Yan et al. (2018) introduced graph networks first proposed by Kipf; Welling (2016) for pose-based activity recognition. Shi et al. (2019) uses a two-stream directed graph neural network for pure pose-based activity recognition. Thus, in future SEU, TEU and intelligent pooling methods through clustering (example FV) could be incorporated into graph-based neural networks for better than state-of-the-art performance.

The results in Table 8.3 further demonstrate the effectiveness of the proposed SEU, TEU (Chapter 7) and (Chapter 7) FV-based activity-aware pooling (Chapter 6) The impact of SEU and TEU and the two-stream architecture in this model is similar to that of the previous chapter (Chapter 7). All the components (two-stream spatial-temporal architecture, the SEU and the TEU) contribute to the model performance in a similar manner in both the models (Sec 7.4.3 and 8.5.1). However, the impact of FV-based activity-aware pooling mechanism on the model presented in this Chapter is significantly more than in both Chapter 6 and NetFV (Miech et al., 2017). The contribution of FV in Chapter 6 and NetFV (Miech et al., 2017) is 0.6% and 1%, respectively. In contrast, in the current model FV-based activity-aware pooling contributes around 6% (Table 8.3) for multi-label and 10% (Table 8.4) for single-label classification. The performance benefit of the FV-based activity aware pooling in this model on NTU-RGBD (Shahroudy et al., 2016) is around 3% which is less in comparison to the multi-label dataset (Chapter 5), but still better than the video-based model introduced in Chapter 6 and NetFV (Miech et al., 2017). The enhanced impact of FV in the current pose-based model could be attributed to the two-stream architecture. The model in Chapter 6 is a single-stream model and although Miech et al. (2017) present a two-stream model, the streams handle different data (video and audio). This means that each type of data (video and audio) is effectively processed by a one stream only. Also, in the current model, FV did not have any impact when used on the base TCN-ResNet (Kim; Reiter, 2017) i.e., without the two-stream architecture. In contrast, in the current two-stream spatio-temporal model, the same data is presented to FV as different representations in the spatial and temporal streams. This allows the FV-based clustering mechanism to learn different aspects (spatial and temporal) of the same data for enanching discriminability. This is also evident from the t-SNE analysis (Sec 8.5.2), which shows that the aggregation (concatenation) of FV from the spatial and temporal stream shows better DBI score than spatial or temporal stream alone.

## 8.7 Conclusion

The main aim of this study is to introduce a method for functional assessment of ADL and this Chapter fulfils this aim (Chapter 1, Sec 1.3). The multi-label activity recognition model presented in this Chapter is able to recognise regular ADL and four different impairment-specific versions of the same ADL when trained on the multi-label dataset presented in Chapter 5. The model takes advantage of the SEU, TEU (Chapter 7) and FV-based activity-aware pooling (Chapter 6) presented previously. The spatial-temporal pose-based model is able to comprehensibly outperform the base TCN-ResNet model and provide close state-of-the-art performance given the constraints of random initialisation and end-to-end trainability. The chapter also presents an elaborate ablation study and analysis of the model to help understand the impact of different aspects on the overall model performance. Chapter 3 to the current Chapter have addressed the main aim and all the five objectives of this research. The next Chapter discusses the contribution of this study to the areas of CV-based rehabilitation and assessment and AI. The model, along with the dataset presented in Chapter 5 has been submitted to the *IEEE International Conference on IROS, 2021.*

# Chapter 9

# Contribution & Conclusion

## 9.1 Introduction

This is the final Chapter of this thesis elaborates the contributions of this work. As shown in Figure 1.2 (Chapter 1), this research focuses on the areas of Health and Social Care, CV and AI. The intended application area of this research is Health and Social Care where DL-based CV methods have been used to advance the research on functional ADL assessment for physically impaired persons. The CV models proposed in this research are based on DL which is sub-area within AI. The Chapter first elaborates the contribution of this research in terms of the stated aim and objectives. Then, the contributions of the models presented in this research to AI is elaborated. As discussed in Chapter 1 (Sec. 1.4) contributing to DL or AI is the core focus of this research. This is followed by a section on the limitations of this work and its reproducibility. The thesis concludes with general recommendations for future research work in the domain of CV-based assessment and rehabilitation of physically impaired persons.

## 9.2 Aim and Objectives: Contribution

The research question posed in Chapter 1 was: *How can a machine or computer recognise different activities of daily living and their variations when executed by healthy individual versus people with different impairments?* This led to the formation of the main aim as:

**Aim**: **The main aim of the research is to contribute a novel model that can not only recognise an ADL, but also discriminate the impairment-specific variations of the same ADL as executed by persons with different physical impairments in comparison to healthy individuals.**

To this end, the study first prepared a dataset that contains 10 different ADL including a healthy and four different impairment-specific version of the same ADL (Chapter 5). Then, this dataset was used to train a multi-label activity recognition model presented in Chapter 8. The model is able to recognise different ADL and their variations when executed by healthy individual and people with different impairments. This answers the main research question and fulfils the main aim of this study. Research on CV-based rehabilitation and assessment is yet to focus on the functional assessment of physically impaired persons through ADL. The review on CV-based methods for rehabilitation and assessment (Chapter 2) shows that authors have approached the topic in several ways such as posture-

recognition, rehabilitation exercise recognition and so on. However, functional assessment of patients through ADL is yet to be fully explored. As explained in Chapter 1 (Sec 1.1.2), ADL recognition is often used to evaluate persons with various physical impairments. Thus, recognising physical impairment specific version of ADL could be the first step towards automated CV-based assessment of physical impaired persons through ADL. Recognising physical impairment-specific version of daily activities of ADL is a major contribution of this research. The next discussion focuses on the contribution of each of the stated objectives:

### 9.2.1 Objectives

**1 To conduct an in-depth and critical review of existing literature in CV-based physical rehabilitation and assessment.**

Chapter 2 reviews the existing literature on CV-based rehabilitation and assessment of physically impaired persons. As discussed in Chapter 1 (Sec. 1.3) and Chapter 2 (Sec. 2.1.1), this is the first study that reviews the recent literature from a CV perspective. It summarises and analyses the CV-based feature extraction and comparison algorithms used by the authors. It also highlights that the mainstream CV community is yet to fully explore this domain. This could be due to the lack of publicly available datasets and DL-based methods which are otherwise ubiquitous in other CV applications. As highlighted in Chapter 2 (Sec. 2.1.1), existing surveys and reviews capture research in this area from a clinical perspective where the focus is on patient rehabilitation, experiment formulation and so on. Thus, the review presented in this research bridges the gap left by the existing current reviews between clinical aspects and CV aspects of this inter-disciplinary domain. The review has been accepted for publication in *Springer Multimedia Systems*.

**2 Make advancement towards lightweight human pose estimation, which is could be used for mobile-based human activity recognition.**

Chapter 2 shows that accurate human body-pose estimation is essential for automated CV-based assessment of physically impaired persons. Moreover, for home or clinic-based pose-estimation where powerful GPUs are often infeasible, lightweight pose estimation is desired. The literature review (Chapter 3.3) shows that there is a gap in existing literature due to unavailability of models for lightweight pose estimation. In this research, the well-known mobile-based DL architecture MobileNets (Howard et al., 2017) is adapted to contribute towards lightweight human pose estimation. This study was presented and published at the $15^{th}$ *IEEE International Conference on AVSS, 2018*.

**3 Prepare a dataset that captures ADL as performed by physically impaired persons.**

The study presents a new dataset that illustrates the difference between an ADL performed by healthy individuals and the impairment-specific variations of the same ADL The size of proposed multi-modal dataset with 5685 videos is well-suited to train contemporary data driven DL-based models. As explained in the literature review (Chapter 3, 3.5 and Chapter 2, 2.9), existing datasets are not suitable to train and evaluate multi-label activity recognition models that need to recognise impairment-specific version of ADL. The dataset is a key contribution of this research and will be made publicly available upon the completion of this study to benefit further research in this area.

**4 Use the latest advancement in the field of DL to develop a novel ADL recognition model.**

The research contributes two novel human activity recognition models. The first model is a purely RGB video-based, introduces a novel FV-based learn-able pooling mechanism (Chapter 6). This model has been accepted for presentation at the *IEEE ICRA, 2021*. The second human activity recognition model demonstrates an effective combination of RGB and human body-pose data and introduces a novel joint position encoding algorithm (Chapter 6). This model has been published at the *IEEE ICPR 2021*. As discussed in Chapter 2, human activity recognition has been used for CV-based assessment and rehabilitation of physically impaired persons. In addition to that, human activity recognition is a widely researched topic in CV owing to its potential applications in wide range of areas including, but not limited to Healthcare (Wang et al., 2013a) and Robotics (Coşar; Bellotto, 2020). The novel activity recognition models have the potential to further advance the research in this area.

**5 Further advance the ADL recognition model to discriminate between different executions of same the ADL.**

The research contributes a new lightweight purely pose-based model which is trained on the multi-label dataset (Chapter 5) to recognise an ADL and its physical impairment-specific variations. The model takes advantage of the joint position-encoding algorithm (Chapter 7) and the learn-able pooling method (Chapter 6) to comprehensively enhance the performance of the TCN-ResNet (Kim; Reiter, 2017) model used as the baseline. Here, the contribution is a multi-label activity recognition model which is able to discriminate subtle intra-class variations within the same activity. Multi-class multi-label (i.e., each sample having multiple labels) image classification has been extensively explored (Wang et al., 2016), but the same is yet to be fully explored for ADL recognition. This model will further the research on multi-class multi-label activity recognition which is useful in situations like CV-based assessment of physically impaired individuals through ADL. The model and the dataset (Chapter 5) has been submitted to the *IEEE International Conference on IROS 2021*.

## 9.3 Contribution to AI

The research presents four different AI models based on DL. In addition to the objective contributions stated in the previous section, each of these models have novel contributions in DL which are discussed next:

### 9.3.1 Lightweight human pose estimation

The model successfully adapts MobileNets (Howard et al., 2017), which is a well-known mobile-based object detection model for a lightweight human pose estimation model. Inspired by the highly successful and widely used stacked hourglass network (Newell et al., 2016), MobileNets is adapted to a hourglass-like architecture which enables the network to be supervised through heat-maps increasing the model performance. In addition, a novel 'Split-Stream' architecture is proposed at the final two layers of the MobileNets which reduces over-fitting and increases accuracy. The 'Split-Stream'

architecture has the potential to be further explored as a more effective alternative to the GAP+FC layer present towards the end of many CNN DL models.

### 9.3.2 Human activity recognition: Model 1

The first model (Chapter 6) introduces a novel learn-able FV with activity-aware pooling mechanism that learns structural information from hidden states of a Bi-LSTM to discriminate the subtle changes in videos, resulting in improved accuracy. To the best of my knowledge, this is the first model to exploit the information contained in hidden attention-focused Bi-LSTM states by semantically clustering them through FV-based learn-able pooling. Also, unlike other learn-able pooling mechanisms (Miech et al., 2017), the method uses activity-aware pooling which obviates the need for further processing through FC layers. The model produces better than state-of-the-art results on the challenging NTU-RGBD dataset with monocular video data. This model highlights the potential of extracting more meaningful information form LSTM. It also has the potential to advance the research on semantic clustering integrated within a DL network.

### 9.3.3 Human activity recognition: Model 2

This model (Chapter 7) introduces a novel human-body pose encoding method that learns the structural relationships and dependencies between various body joints, as well as captures long-term temporal dependencies of each body joint. As explained in the literature review (Chapter 3, Sec 3.4.3), authors have used various hand-crafted mechanisms to encode human body-pose to enhance the model performance. In contrast, the proposed method *learns* these encodings for a more effective representation of the structural relationships and dependencies between various body joints. The proposed spatial and temporal encoding implemented through SEU and TEU respectively, is the main contribution of this model. The final 'Attention'-driven model consisting of two pose streams (spatial and temporal) and a RGB stream, achieves state-of-the-art results across three datasets including the challenging NTU-RGBD dataset. The model shows the impact of learn-able body-pose encoding and integration of RGB video data with human-body pose data.

### 9.3.4 Functional activity recognition

The multi-label functional human activity recognition model (Chapter 8) takes advantage of the learn-able pooling method and the body-pose encoding method introduced in Chapter 6 and 7 respectively. This gives a lightweight activity recognition model that is purely based on human body-pose data. The model outperforms the base TCN-ResNet (Kim; Reiter, 2017) comprehensibly, thereby showing the effectiveness of the proposed architecture. The model also outperforms other existing state-of-the-art architectures under the constraints of random initialisation and end-to-end trainablity. As discussed in Chapter 8, the model is much lighter than other existing models while delivering comparable performance.

## 9.4 Limitations

This section highlights the limitations of the current work:

1. The lightweight pose estimation model presented in Chapter 4 will need to incorporate full-body and 3D pose-estimation before it can be used for practical applications.
2. The dataset presented in Chapter 5 captures eight different impairments. But, physical impairments manifests in a huge variety of forms and any dataset targeted towards automated assessments of patients will need to capture a wider range of impairments for practical applications.
3. Due to time constraint, the dataset presented in Chapter 5 illustrates a single form of execution for each 'Activity' and 'Impairment' combination. For example, it exhibits 'shoulder weakness' through leaning to a side but, the same may be also exhibited by a limited range of shoulder motion. For practical applications, wider manifestations of such impairments will need to be included in the dataset.
4. Owing to the use of pre-trained Inception-Resnet-V2 (Szegedy et al., 2017), the human activity recognition models presented in Chapter 6 and 7 has a large number of parameters. Therefore, these models require powerful GPUs for training and inference which may make it difficult for real-time implementations in a home or clinic-based scenario.
5. The pose-based multi-label activity recognition model presented in Chapter 8 has further scope for improvement in performance. As shown in Table 8.1, the model achieves less than state-of-the-art results on the NTU-RGBD (Shahroudy et al., 2016) dataset.

## 9.5 Reproducibility

The thesis presents four DL-based models and the following steps have been taken to ensure the experimental reproducibility of these models:

1. **Model Training:** The hardware (e.g., server used, GPUs, etc.) and software resources (e.g., Tensorflow, Keras) used for training and validating the models are described in the respective Chapters (Chapter 4 Sec. 4.6, Chapter 6 Sec. 6.5, Chapter 7 Sec. 7.4, Chapter 8 Sec. 8.4). These sections also describe the training methodology (example number of epochs, fine-tuning) and the hyper-parameters used.
2. **Datasets:** The performance of DL models depend on the data used to train the model. Although, the goal in this thesis was to train the activity recognition model on the multi-label dataset presented in Chapter 5, the models were evaluated with well-known publicly available datasets to ensure fair comparison to other state-of-the-art models.
3. **Code availability:** The dataset presented in Chapter 5 and the code for the DL models will be made publicly available on completion of this research.

## 9.6 Conclusion

To summarise, the main aim of the study was to contribute towards the domain of CV-based functional assessment of ADL for physically impaired individuals. To this end, the work first prepared a dataset that presents ADL as performed by physically impaired persons. Then the dataset was used to train a new DL-based model to discriminate a perfectly executed ADL from impairment-specific versions of the same. To the best of my knowledge, this is the first study that explores CV-based functional assessment of patient ADL in the form of multi-class multi-label human activity recognition. The study contributes a literature review of CV-based rehabilitation and assessment methods that is representative of the existing literature in this domain. The study also contributes a novel DL-based human pose estimation model and three human activity recognition models. The research contributes towards the field of DL through the 'Split-Stream' architecture (Chapter 4), the FV-based activity-aware pooling mechanism (Chapter 6 and 8) and the novel human body-pose encoding algorithm (Chapter 7 and 8). As discussed in Chapter 2, the CV community is yet to fully explore the area of CV-based rehabilitation and assessment. This is especially true with regards to the application of DL-methods which are extensively used in other CV applications due to their impressive performance. In contrast, DL-based methods have seen comparatively less use in CV-based rehabilitation and assessment. This work, especially the multi-label activity recognition model and the dataset has the potential to attract attention of the CV and AI community towards further research in this area involving DL.

In future, this research could be extended in two major directions. First, for CV-based rehabilitation and assessment of physically impaired persons, a larger dataset could be created that captures more physical impairment conditions. The dataset could also include the severity or the extent of a physical impairment in different patients. Furthermore, the activities and their corresponding impairments could be captured in realistic day to day scenarios (e.g., reaching above in kitchen with tremors). A major goal would be to include actual patients instead of just relying on actors. All these changes will help to create a dataset that can train a AI model for deployment in home or at a clinic. Second, in statistical or intelligent processing, clustering methods are generally used for unsupervised learning (Goodfellow et al., 2016). However, the literature review (Chapter 3, Sec 3.2.3) showed that authors have integrated clustering algorithms in supervised DL-based approaches for intelligent pooling for performance gain over statistical pooling. Researchers have also used semi-supervised learning to learn from a small amount of labelled data in combination with larger amount of unlabelled data to overcome issues arising from lack of labelled data (Reddy et al., 2018). Thus, in future, the current approach involving FV-based clustering could be extended to semi-supervised learning for multi-label human activity recognition. This would be especially helpful in the domain of CV-based rehabilitation and assessment where labelled dataset is a scarcity. With a semi-supervised model one can use random videos available on the internet to train their model. Integrating clustering with DL is an active area of research which is evident from models like Spectral Clustering (Shaham et al., 2018), Deep Embedded Clustering (Ren et al., 2019) and so on. In future, instead of FV, one can explore these methods as an effective alternative to statistical pooling. The literature review (Chapter 3, Sec. 3.4.2) shows that authors have used additional hand-crafted information with body-pose such including velocity, super-normalisation and others to increase ADL recognition accuracy.

In the current study the SEU and TEU 'learns' this information. All these methods effectively introduce prior information to the model to improve performance. In past authors have used template matching for activity recognition. Polana; Nelson (1994b) matched sequences against a spatio-temporal template of motion features for sequence or activity recognition. Seto et al. (2015) used template selection approach based on DTW, so that complex feature extraction or domain knowledge could be avoided. Thus, in future pose-based template information could be incorporated as prior information in a deep network to increase the recognition accuracy. Finally, the thesis concludes with recommendations for future research in CV-based assessment and rehabilitation:

1. Large scale publicly available datasets consisting of physically impaired human movements need to be created for training AI models for practical applications.
2. Creating synthetic data from GAN, unsupervised or semi-supervised learning are some other the techniques that may be used to compliment the datasets and minimise the impact of lack of data.
3. Kinect has its own limitations as shown by Webster; Celik (2014), and thus more recently available devices such as Orbec Astra (Coroiu; Coroiu, 2018) or DL based 3D tracking algorithms can be used (Yang et al., 2018; Pavlakos et al., 2018).
4. The use of image or CV-based features needs to be explored for capturing contextual information. Only using skeletal features leads to loss of information such as optical flow, semantic segmentation, contextual information and so on.
5. More recent techniques such as 'Attention'-based methods, GANs, TCN, ODENets, Graph Neural Networks and so can be explored to encode, learn and compare patient activity with that of healthy individuals.
6. For practical and commercial applicability, more attention from the CV community is required. The lack of attention from mainstream CV community is evident by the absence of articles in this domain in premiere CV conferences and journals.

# References


ADAMS, R. J.; LICHTER, M. D.; KREPKOVICH, E. T.; ELLINGTON, A.; WHITE, M.; DIAMOND, P. T., 2015. Assessing upper extremity motor function in practice of virtual activities of daily living. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. Vol. 23, no. 2, pp. 287–296.

AGGARWAL, J., 2004. Semantic-level Understanding of Human Actions and Interactions using Event Hierarchy. *CVPR Workshop*, pp. 12–12. Available from DOI: 10.1109/CVPR.2004.434.

AHAD, M. A. R.; ANTAR, A. D.; SHAHID, O., 2019. Vision-based Action Understanding for Assistive Healthcare: A Short Review. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops 2019*, pp. 1–11.

AKOPYAN, M.; KHASHBA, E., 2017. Large-Scale YouTube-8M Video Understanding with Deep Neural Networks. Available from arXiv: 1706.04488.

ALIAKBARIAN, M. S.; SALEH, F. S.; SALZMANN, M.; FERNANDO, B.; PETERSSON, L.; ANDERSSON, L., 2017. Encouraging LSTMs to Anticipate Actions Very Early. *ICCV*. Available from arXiv: 1703.07023.

ALOM, M. Z.; TAHA, T. M.; YAKOPCIC, C.; WESTBERG, S.; SIDIKE, P.; NASRIN, M. S.; VAN ESESN, B. C.; AWWAL, A. A. S.; ASARI, V. K., 2018. The history began from alexnet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164*.

ALOYSIUS, N.; GEETHA, M., 2017. A review on deep convolutional neural networks. In: *2017 International Conference on Communication and Signal Processing (ICCSP)*, pp. 0588–0592.

ANTÓN, D.; GOÑI, A.; ILLARRAMENDI, A.; TORRES-UNDA, J. J.; SECO, J., 2013. KiReS: A Kinect-based telerehabilitation system. In: *e-Health Networking, Applications & Services (Healthcom), 2013 IEEE 15th International Conference on*, pp. 444–448.

ANTONIOU, A.; STORKEY, A.; EDWARDS, H., 2017. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*.

ANTUNES, J.; BERNARDINO, A.; SMAILAGIC, A.; SIEWIOREK, D. P., 2018. AHA-3D: A Labelled Dataset for Senior Fitness Exercise Recognition and Segmentation from 3D Skeletal Data. In: *BMVC*, p. 332.

ANTUNES, M.; BAPTISTA, R.; DEMISSE, G.; AOUADA, D.; OTTERSTEN, B., 2016. Visual and human-interpretable feedback for assisting physical activity. In: *ECCV*, pp. 115–129.

ARAFAH, M.; MOGHLI, Q. A., 2016. Efficient image recognition technique using invariant moments and principle component analysis. *Journal of Data Analysis and Information Processing*. Vol. 5, no. 1, pp. 1–10.

ARANDJELOVIC, R.; GRONAT, P.; TORII, A.; PAJDLA, T.; SIVIC, J., 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In: *in Proc. of the CVPR*, pp. 5297–5307.

ARJOVSKY, M.; CHINTALA, S.; BOTTOU, L., 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

ASA, 2019. *Hemiparesis*. American Stroke Association. Available also from: `https://www.stroke.org/en/about-stroke/effects-of-stroke/physical-effects-of-stroke/physical-impact/hemiparesis`.

AVILÉS, H.; LUIS, R.; OROPEZA, J.; ORIHUELA-ESPINA, F.; LEDER, R.; HERNÁNDEZ-FRANCO, J.; SUCAR, E., 2011. Gesture Therapy 2.0: Adapting the rehabilitation therapy to the patient progress. *Probabilistic Problem Solving in BioMedicine*, p. 3.

AVOLA, D.; CINQUE, L.; FORESTI, G. L.; MARINI, M. R.; PANNONE, D., 2018. VRheab: a fully immersive motor rehabilitation system based on recurrent neural network. *Multimedia Tools and Applications*, pp. 1–28.

BA, J. L.; KIROS, J. R.; HINTON, G. E., 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

BACCOUCHE, M.; MAMALET, F.; WOLF, C.; GARCIA, C.; BASKURT, A., 2010. Action classification in soccer videos with long short-term memory recurrent neural networks. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer, Berlin, Heidelberg. Vol. 6353 LNCS, pp. 154–159. ISBN 3642158218. ISSN 03029743. Available from DOI: `10.1007/978-3-642-15822-3_20`.

BADRINARAYANAN, V.; KENDALL, A.; CIPOLLA, R., 2017. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*. Vol. 39, no. 12, pp. 2481–2495.

BAHDANAU, D.; CHO, K.; BENGIO, Y., 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

BALDI, P., 2012. Autoencoders, unsupervised learning, and deep architectures. In: *Proceedings of ICML workshop on unsupervised and transfer learning*, pp. 37–49.

BANFIELD, R. E.; HALL, L. O.; BOWYER, K. W.; KEGELMEYER, W. P., 2006. A comparison of decision tree ensemble creation techniques. *IEEE transactions on pattern analysis and machine intelligence*. Vol. 29, no. 1, pp. 173–180.

BAPTISTA, R.; GHORBEL, E.; SHABAYEK, A. E. R.; MOISSENET, F.; AOUADA, D.; DOUCHET, A.; ANDRÉ, M.; PAGER, J.; BOUILLAND, S., 2019. Home self-training: Visual feedback for assisting physical activity for stroke survivors. *Computer methods and programs in biomedicine*. Vol. 176, pp. 111–120.

BAPTISTA, R.; GONCALVES ALMEIDA ANTUNES, M.; AOUADA, D.; OTTERSTEN, B., 2017. Video-based feedback for assisting physical activity. In: *VISAPP*.

BARADEL, F.; WOLF, C.; MILLE, J., 2017. Human action recognition: Pose-based attention draws focus to hands. In: *in Proc. of the ICCV*, pp. 604–613.

BARADEL, F.; WOLF, C.; MILLE, J., 2018a. Human activity recognition with pose-driven attention to rgb. In: *in Proc. of the BMVC*.

BARADEL, F.; WOLF, C.; MILLE, J.; TAYLOR, G. W., 2018b. Glimpse Clouds: Human Activity Recognition From Unstructured Feature Points. In: *in Proc. of the CVPR*.

BATRA, D.; CHEN, T.; SUKTHANKAR, R., 2008. Space-time shapelets for action recognition. In: *2008 IEEE Workshop on Motion and Video Computing, WMVC*. IEEE, pp. 1–6. ISBN 1424420008. Available from DOI: 10.1109/WMVC.2008.4544051.

BAUMGARTNER, R. N.; KOEHLER, K. M.; GALLAGHER, D.; ROMERO, L.; HEYMSFIELD, S. B.; ROSS, R. R.; GARRY, P. J.; LINDEMAN, R. D., 1998. Epidemiology of sarcopenia among the elderly in New Mexico. *American journal of epidemiology*. Vol. 147, no. 8, pp. 755–763.

BAY, H.; TUYTELAARS, T.; VAN GOOL, L., 2006. Surf: Speeded up robust features. In: *ECCV*, pp. 404–417.

BEHERA, A.; COHN, A. G.; HOGG, D. C., 2014. Real-time activity recognition by discerning qualitative relationships between randomly chosen visual features. In: *BMVC 2014-Proceedings of the British Machine Vision Conference 2014*.

BEHERA, A.; KEIDEL, A.; DEBNATH, B., 2018. Context-driven multi-stream LSTM (M-LSTM) for recognizing fine-grained activity of drivers. In: *German Conference on Pattern Recognition*, pp. 298–314.

BENETTAZZO, F.; IARLORI, S.; FERRACUTI, F.; GIANTOMASSI, A.; ORTENZI, D.; FREDDI, A.; MONTERIÙ, A.; INNOCENZI, S.; CAPECCI, M.; CERAVOLO, M. G., et al., 2015. Low cost RGB-D vision based system to support motor disabilities rehabilitation at home. In: *Ambient Assisted Living*. Springer, pp. 449–461.

BESL, P. J.; JAIN, R. C., 1985. Invariant surface characteristics for 3-D object recognition in range images. *Computer Vision, Graphics, and Image Processing*. Vol. 31, no. 3, p. 400.

BIGONI, M.; BAUDO, S.; CIMOLIN, V.; CAU, N.; GALLI, M.; PIANTA, L.; TACCHINI, E.; CAPODAGLIO, P.; MAURO, A., 2016. Does kinematics add meaningful information to clinical assessment in post-stroke upper limb rehabilitation? A case report. *Journal of physical therapy science*. Vol. 28, no. 8, pp. 2408–2413.

BISHOP, C. M., 2006. *Pattern recognition and machine learning*. springer.

BRADSKI, G. R.; DAVIS, J. W., 2002. Motion segmentation and pose recognition with motion history gradients. *Machine Vision and Applications*. Vol. 13, no. 3, pp. 174–184.

BREGLER, C., 1997. Learning and recognizing human dynamics in video sequences. In: *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 568–574.

BRUHN, A.; WEICKERT, J.; SCHNÖRR, C., 2005. Lucas/Kanade meets Horn/Schunck: Combining local and global optic flow methods. *International journal of computer vision*. Vol. 61, no. 3, pp. 211–231.

BRUNA, J.; ZAREMBA, W.; SZLAM, A.; LECUN, Y., 2013. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*.

CAMEIRÃO, M. S.; BADIA, S. B. i; OLLER, E. D.; VERSCHURE, P. F., 2010. Neurorehabilitation using the virtual reality based Rehabilitation Gaming System: methodology, design, psychometrics, usability and validation. *Journal of neuroengineering and rehabilitation*. Vol. 7, no. 1, p. 48.

CAO, Z.; HIDALGO, G.; SIMON, T.; WEI, S.-E.; SHEIKH, Y., 2018. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In: *arXiv preprint arXiv:1812.08008*.

CAO, Z.; SIMON, T.; WEI, S.-E.; SHEIKH, Y., 2016a. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. Available from DOI: 10.1109/CVPR.2017.143.

CAO, Z.; SIMON, T.; WEI, S.-E.; SHEIKH, Y., 2016b. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. Available from DOI: 10.1109/CVPR.2017.143.

CAO, Z.; SIMON, T.; WEI, S.-E.; SHEIKH, Y., 2017. Realtime multi-person 2d pose estimation using part affinity fields. In: *CVPR*. Vol. 1, p. 7.

CAPECCI, M.; CERAVOLO, M. G.; FERRACUTI, F.; IARLORI, S.; KYRKI, V.; LONGHI, S.; ROMEO, L.; VERDINI, F., 2016. Physical rehabilitation exercises assessment based on hidden semi-markov model by kinect v2. In: *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pp. 256–259.

CAPECCI, M.; CERAVOLO, M. G.; FERRACUTI, F.; IARLORI, S.; KYRKI, V.; MONTERIÙ, A.; ROMEO, L.; VERDINI, F., 2018. A Hidden Semi-Markov Model based approach for rehabilitation exercise assessment. *Journal of biomedical informatics*. Vol. 78, pp. 1–11.

CAPECCI, M.; CERAVOLO, M. G.; FERRACUTI, F.; IARLORI, S.; MONTERIÙ, A.; ROMEO, L.; VERDINI, F., 2019. The KIMORE Dataset: KInematic Assessment of MOvement and Clinical Scores for Remote Monitoring of Physical REhabilitation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*. Vol. 27, no. 7, pp. 1436–1448.

CARREIRA, J.; ZISSERMAN, A., 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In: *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308.

CARY, F.; POSTOLACHE, O.; GIRAO, P. S., 2014. Kinect based system and artificial neural networks classifiers for physiotherapy assessment. In: *Medical Measurements and Applications (MeMeA), 2014 IEEE International Symposium on*, pp. 1–6.

CHANG, C.-Y.; LANGE, B.; ZHANG, M.; KOENIG, S.; REQUEJO, P.; SOMBOON, N.; SAWCHUK, A. A.; RIZZO, A. A., et al., 2012. Towards pervasive physical rehabilitation using Microsoft Kinect. In: *PervasiveHealth*, pp. 159–162.

CHANG, Y.-J.; CHEN, S.-F.; HUANG, J.-D., 2011. A Kinect-based system for physical rehabilitation: A pilot study for young adults with motor disabilities. *Research in developmental disabilities*. Vol. 32, no. 6, pp. 2566–2570.

CHANG, Y.-J.; HAN, W.-Y.; TSAI, Y.-C., 2013. A Kinect-based upper limb rehabilitation system to assist people with cerebral palsy. *Research in developmental disabilities*. Vol. 34, no. 11, pp. 3654–3659.

CHEN, C.; JAFARI, R.; KEHTARNAVAZ, N., 2015. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In: *ICIP*, pp. 168–172.

CHEN, C.-H.; RAMANAN, D., 2017. 3d human pose estimation= 2d pose estimation+ matching. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7035–7043.

CHEN, H.; WANG, G.; XUE, J.-H.; HE, L., 2016. A novel hierarchical framework for human action recognition. *Pattern Recognition*. Vol. 55, pp. 148–159. ISBN 0031-3203. ISSN 00313203. Available from DOI: 10.1016/j.patcog.2016.01.020.

CHEN, T. Q.; RUBANOVA, Y.; BETTENCOURT, J.; DUVENAUD, D. K., 2018a. Neural ordinary differential equations. In: *Advances in neural information processing systems*, pp. 6571–6583.

CHEN, X.; YUILLE, A. L., 2014. Articulated pose estimation by a graphical model with image dependent pairwise relations. In: *NIPS*, pp. 1736–1744.

CHEN, Y.-L.; LIU, C.-H.; YU, C.-W.; LEE, P.; KUO, Y.-W., 2018b. An Upper Extremity Rehabilitation System Using Efficient Vision-Based Action Identification Techniques. *Applied Sciences*. Vol. 8, no. 7, p. 1161.

CHENG, J.; DONG, L.; LAPATA, M., 2016. Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*.

CHO, C.-W.; CHAO, W.-H.; LIN, S.-H.; CHEN, Y.-Y., 2009. A vision-based analysis system for gait recognition in patients with Parkinson's disease. *Expert Systems with applications*. Vol. 36, no. 3, pp. 7033–7039.

CHO, K.; COURVILLE, A.; BENGIO, Y., 2015. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Transactions on Multimedia*. Vol. 17, no. 11, pp. 1875–1886.

CHO, K.; VAN MERRIËNBOER, B.; BAHDANAU, D.; BENGIO, Y., 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

CHOLLET, F., 2017. Xception: Deep learning with depthwise separable convolutions, pp. 1251–1258.

CHOMAT, O.; CROWLEY, J. L., 1998. Recognizing motion using local appearance. In: *nternational Symposium on Intelligent Robotic Systems, University of Edinburgh*. Available also from: `https://pdfs.semanticscholar.org/aa00/ee8bb59d31aabeaf6b407dbd653a17e2a4ed.pdf`.

CHRISTENSEN, R., 2018. *Analysis of variance, design, and regression: Linear modeling for unbalanced data*. CRC Press.

CHU, W.-S.; ZHOU, F.; DE LA TORRE, F., 2012. Unsupervised temporal commonality discovery. In: *ECCV*, pp. 373–387.

CIABATTONI, L.; FERRACUTI, F.; IARLORI, S.; LONGHI, S.; ROMEO, L., 2016a. A novel computer vision based e-rehabilitation system: From gaming to therapy support. In: *Consumer Electronics (ICCE), 2016 IEEE International Conference on*, pp. 43–44.

CIABATTONI, L.; FERRACUTI, F.; LAZZARO, G.; ROMEO, L.; VERDINI, F., 2016b. Serious gaming approach for physical activity monitoring: A visual feedback based on quantitative evaluation. In: *International Conference on Consumer Electronics*, pp. 209–213.

COCO, 2016. *MSCOCO keypoint challenge 2016* [online] [visited on 2021-05-08]. Available from: `https://cocodataset.org/#home`.

COIFMAN, R. R.; LAFON, S., 2006. Diffusion maps. *Applied and computational harmonic analysis*. Vol. 21, no. 1, pp. 5–30.

COMPTROLLER AND AUDITOR GENERAL, 2015. Services for people with neurological conditions: progress review | National Audit Office (NAO). *National Audit Office*. No. July, para 1.19. ISBN 9781904219958. Available also from: `https://www.nao.org.uk/report/services-for-people-with-neurological-conditions-progress-review/`.

COROIU, A. D. C. A.; COROIU, A., 2018. Interchangeability of Kinect and Orbbec sensors for gesture recognition. In: *2018 IEEE 14th international conference on intelligent computer communication and processing (ICCP)*, pp. 309–315.

COŞAR, S.; BELLOTTO, N., 2020. Human Re-Identification with a Robot Thermal Camera Using Entropy-Based Sampling. *Journal of Intelligent & Robotic Systems*. Vol. 98, no. 1, pp. 85–102.

CUELLAR, M. P.; ROS, M.; MARTIN-BAUTISTA, M. J.; LE BORGNE, Y.; BONTEMPI, G., 2014. An approach for the evaluation of human activities in physical therapy scenarios. In: *International Conference on Mobile Networks and Management*, pp. 401–414.

DA GAMA, A.; CHAVES, T.; FIGUEIREDO, L.; TEICHRIEB, V., 2012. Guidance and movement correction based on therapeutics movements for motor rehabilitation support systems. In: *2012 14th Symposium on Virtual and Augmented Reality*, pp. 191–200.

DA GAMA, A.; FALLAVOLLITA, P.; TEICHRIEB, V.; NAVAB, N., 2015a. Motor Rehabilitation Using Kinect: A Systematic Review. *Games for Health Journal*. Vol. 4, no. 2, pp. 123–135. ISBN 2161-783X. ISSN 2161-783X. Available from DOI: 10.1089/g4h.2014.0047.

DA GAMA, A.; FALLAVOLLITA, P.; TEICHRIEB, V.; NAVAB, N., 2015b. Motor rehabilitation using Kinect: a systematic review. *Games for health journal*. Vol. 4, no. 2, pp. 123–135.

DALAL, N.; TRIGGS, B.; SCHMID, C., 2006. Human detection using oriented histograms of flow and appearance. In: *European conference on computer vision*, pp. 428–441.

DAS, S.; DAI, R.; KOPERSKI, M.; MINCIULLO, L.; GARATTONI, L.; BREMOND, F.; FRANCESCA, G., 2019. Toyota smarthome: Real-world activities of daily living. In: *Proceedings of the ICCV*, pp. 833–842.

DAVIES, D. L.; BOULDIN, D. W., 1979. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*. No. 2, pp. 224–227.

DAVIS, J. W.; BOBICK, A. F., 1997. The representation and recognition of human movement using temporal templates. In: *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 928–934.

DEMISSE, G. G.; PAPADOPOULOS, K.; AOUADA, D.; OTTERSTEN, B., 2018. Pose encoding for robust skeleton-based action recognition. In: *in Proc. of the CVPR*, pp. 188–194.

DENG, Z.; VAHDAT, A.; HU, H.; MORI, G., 2016. Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition. In: *in Proc. of the ICCV*, pp. 4772–4781.

DESAI, K.; BAHIRAT, K.; RAMALINGAM, S.; PRABHAKARAN, B.; ANNASWAMY, T.; MAKRIS, U. E., 2016. Augmented reality-based exergames for rehabilitation. In: *Proceedings of the 7th International Conference on Multimedia Systems*, p. 22.

DEVANNE, M. et al., 2018a. Generating Shared Latent Variables for Robots to Imitate Human Movements and Understand their Physical Limitations. In: *ECCV*, pp. 190–197.

DEVANNE, M.; REMY-NERIS, O.; LE GALS-GARNETT, B.; KERMARREC, G.; THEPAUT, A., et al., 2018b. A co-design approach for a rehabilitation robot coach for physical rehabilitation based on the error classification of motion errors. In: *2018 Second IEEE International Conference on Robotic Computing (IRC)*, pp. 352–357.

DOLATABADI, E.; ZHI, Y. X.; YE, B.; COAHRAN, M.; LUPINACCI, G.; MIHAILIDIS, A.; WANG, R.; TAATI, B., 2017. The toronto rehab stroke pose dataset to detect compensation during stroke rehabilitation therapy. In: *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pp. 375–381.

DONAHUE, J.; HENDRICKS, L. A.; ROHRBACH, M.; VENUGOPALAN, S.; GUADARRAMA, S.; SAENKO, K.; DARRELL, T., 2016. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 39, no. 4, pp. 677–691. ISBN 9781467369640. ISSN 01628828. Available from DOI: 10.1109/TPAMI.2016.2599174.

DU, Y.; WANG, W.; WANG, L., 2015. Hierarchical recurrent neural network for skeleton based action recognition. In: *in Proc. of the CVPR*, pp. 1110–1118.

DYSHEL, M.; ARKADIR, D.; BERGMAN, H.; WEINSHALL, D., 2015. Quantifying Levodopa-Induced Dyskinesia Using Depth Camera. In: *ICCV Workshops*, pp. 119–126.

EDEMEKONG PF, e. a., 2020. Activities of Daily Living. Available also from: https://www.ncbi.nlm.nih.gov/books/NBK470404/.

EICHLER, N.; HEL-OR, H.; SHMISHONI, I.; ITAH, D.; GROSS, B.; RAZ, S., 2018. Non-invasive motion analysis for stroke rehabilitation using off the shelf 3d sensors. In: *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8.

EINARSSON, G.; CLEMMENSEN, L. K.; RUDÅ, D.; FINK-JENSEN, A.; NIELSEN, J. B.; PAGSBERG, A. K.; WINGE, K.; PAULSEN, R. R., 2018. Computer Aided Identification of Motion Disturbances Related to Parkinson's Disease. In: *International Workshop on PRedictive Intelligence In MEdicine*, pp. 1–8.

ESTEBAN, C.; HYLAND, S. L.; RÄTSCH, G., 2017. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633*.

ESTEVA, A.; ROBICQUET, A.; RAMSUNDAR, B.; KULESHOV, V.; DEPRISTO, M.; CHOU, K.; CUI, C.; CORRADO, G.; THRUN, S.; DEAN, J., 2019. A guide to deep learning in healthcare. *Nature medicine*. Vol. 25, no. 1, pp. 24–29.

EWEIWI, A.; CHEEMA, M. S.; BAUCKHAGE, C.; GALL, J., 2014. Efficient pose-based action recognition. In: *in Proc. of the ACCV*, pp. 428–443.

EXELL, T.; FREEMAN, C.; MEADMORE, K.; KUTLU, M.; ROGERS, E.; HUGHES, A.-M.; HALLEWELL, E.; BURRIDGE, J., 2013. Goal orientated stroke rehabilitation utilising electrical stimulation, iterative learning and Microsoft Kinect. In: *Rehabilitation robotics (icorr), 2013 ieee international conference on*, pp. 1–6.

FANG, H.; XIE, S.; TAI, Y.-W.; LU, C., 2017. Rmpe: Regional multi-person pose estimation. In: *ICCV*. Vol. 2.

FARHA, Y. A.; GALL, J., 2019. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3575–3584.

FARNEBÄCK, G., 2003. Two-frame motion estimation based on polynomial expansion. In: *Scandinavian conference on Image analysis*, pp. 363–370.

FASOLA, J.; MATARIĆ, M. J., 2013. A socially assistive robot exercise coach for the elderly. *Journal of Human-Robot Interaction*. Vol. 2, no. 2, pp. 3–32.

FERN'NDEZ-BAENA, A.; SUSIN, A.; LLIGADAS, X., 2012. Biomechanical validation of upper-body and lower-body joint movements of kinect motion capture data for rehabilitation treatments. In: *Intelligent networking and collaborative systems (INCoS), 2012 4th international conference on*, pp. 656–661.

FERNANDO, B.; GAVVES, E.; ORAMAS, J.; GHODRATI, A.; TUYTELAARS, T., 2016. Rank pooling for action recognition. *IEEE transactions on PAMI*. Vol. 39, no. 4, pp. 773–787.

FERRUCCI, L.; KOH, C.; BANDINELLI, S.; GURALNIK, J., 2010. Disability, functional status, and activities of daily living. In: *Encyclopedia of gerontology*. Elsevier Inc., pp. 427–436.

FISCHLER, M. A.; BOLLES, R. C., 1981. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*. Vol. 24, no. 6, pp. 381–395.

FORSYTH, D. A.; PONCE, J., 2012. *Computer vision: a modern approach*. Prentice Hall Professional Technical Reference.

FREUND, Y.; SCHAPIRE, R.; ABE, N., 1999. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*. Vol. 14, no. 771-780, p. 1612.

FRISOLI, A.; LOCONSOLE, C.; LEONARDIS, D.; BANNO, F.; BARSOTTI, M.; CHISARI, C.; BERGAMASCO, M., 2012. A new gaze-BCI-driven control of an upper limb exoskeleton for rehabilitation in real-world tasks. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*. Vol. 42, no. 6, pp. 1169–1179.

FU, Y.; WANG, X.; ZHU, Z.; TAN, J.; ZHAO, Y.; DING, Y.; CHEN, W., 2020. Vision-based Automatic Detection of Compensatory Postures of after-Stroke Patients During Upper-extremity Robot-assisted Rehabilitation: A Pilot Study in Reaching Movement. In: *2020 International Conference on Assistive and Rehabilitation Technologies (iCareTech)*, pp. 62–66.

GALEANO, D.; BRUNETTI, F.; TORRICELLI, D.; PIAZZA, S.; PONS, J. L., 2014. A tool for balance control training using muscle synergies and multimodal interfaces. *BioMed research international*. Vol. 2014.

GALNA, B.; BARRY, G.; JACKSON, D.; MHIRIPIRI, D.; OLIVIER, P.; ROCHESTER, L., 2014. Accuracy of the Microsoft Kinect sensor for measuring movement in people with Parkinson's disease. *Gait & posture*. Vol. 39, no. 4, pp. 1062–1068.

GAUTHIER, J., 2014. Conditional generative adversarial nets for convolutional face generation. *Class Project for Stanford CS231N: Convolutional Neural Networks for Visual Recognition, Winter semester*. Vol. 2014, no. 5, p. 2.

GERBER, S.; TASDIZEN, T.; WHITAKER, R., 2007. Robust non-linear dimensionality reduction using successive 1-dimensional Laplacian eigenmaps. In: *Proceedings of the 24th international conference on Machine learning*, pp. 281–288.

GHALI, A.; CUNNINGHAM, A. S.; PRIDMORE, T. P., 2003. Object and event recognition for stroke rehabilitation. In: *Visual Communications and Image Processing 2003*. Vol. 5150, pp. 980–990.

GIRDHAR, R.; RAMANAN, D., 2017. Attentional pooling for action recognition. In: *in Proc. of the NIPS*, pp. 34–45.

GIRDHAR, R.; RAMANAN, D.; GUPTA, A.; SIVIC, J.; RUSSELL, B., 2017. Actionvlad: Learning spatio-temporal aggregation for action classification. In: *in Proc. of the CVPR*, pp. 971–980.

GKIOXARI, G.; HARIHARAN, B.; GIRSHICK, R.; MALIK, J., 2014. R-CNNs for Pose Estimation and Action Detection. Available from arXiv: `1406.5212`.

GLADSTONE, D. J.; DANELLS, C. J.; BLACK, S. E., 2002. The Fugl-Meyer assessment of motor recovery after stroke: a critical review of its measurement properties. *Neurorehabilitation and neural repair*. Vol. 16, no. 3, pp. 232–240.

GOFFREDO, M.; SCHMID, M.; CONFORTO, S.; CARLI, M.; NERI, A.; D'ALESSIO, T., 2009. Markerless human motion analysis in Gauss–Laguerre transform domain: An application to sit-to-stand in young and elderly people. *IEEE Transactions on Information Technology in Biomedicine*. Vol. 13, no. 2, pp. 207–216.

GONG, D.; MEDIONI, G., 2011. Dynamic manifold warping for view invariant action recognition. In: *2011 International Conference on Computer Vision*, pp. 571–578.

GONZÁLEZ, A.; HAYASHIBE, M.; FRAISSE, P., 2012. Three dimensional visualization of the statically equivalent serial chain from kinect recording. In: *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pp. 4843–4846.

GONZÁLEZ-ORTEGA, D.; DIAZ-PERNAS, F.; MARTINEZ-ZARZUELA, M.; ANTÓN-RODRIGUEZ, M., 2014. A Kinect-based system for cognitive rehabilitation exercises monitoring. *Computer methods and programs in biomedicine*. Vol. 113, no. 2, pp. 620–631.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A., 2016. *Deep Learning*. MIT Press. Available also from: `http://www.deeplearningbook.org`.

GORELICK, L.; BLANK, M.; SHECHTMAN, E.; IRANI, M.; BASRI, R., 2007. Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Vol. 29, no. 12, pp. 2247–2253. ISBN 076952334X. ISSN 01628828. Available from DOI: `10.1109/TPAMI.2007.70711`.

GREEN, J.; YOUNG, J., 2001. A test-retest reliability study of the Barthel Index, the Rivermead Mobility Index, the Nottingham Extended Activities of Daily Living Scale and the Frenchay Activities Index in stroke patients. *Disability and rehabilitation*. Vol. 23, no. 15, pp. 670–676.

GRUSHIN, A.; MONNER, D. D.; REGGIA, J. A.; MISHRA, A., 2013. Robust Human Action Recognition via Long Short-Term Memory. In: *International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–8. ISBN 978-1-4673-6129-3. ISSN 2161-4393. Available from DOI: `10.1109/IJCNN.2013.6706797`.

GU, Y.; PANDIT, S.; SARAEE, E.; NORDAHL, T.; ELLIS, T.; BETKE, M., 2019. Home-based physical therapy with an interactive computer vision system. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0.

HABIBIAN, A.; MENSINK, T.; SNOEK, C. G., 2016. Video2vec embeddings recognize events when examples are scarce. *IEEE transactions on PAMI*. Vol. 39, no. 10, pp. 2089–2103.

HAGIHARA, H.; IENAGA, N.; ENOMOTO, D.; TAKAHATA, S.; ISHIHARA, H.; NODA, H.; TSUDA, K.; TERAYAMA, K., 2020. Computer Vision-Based Approach for Quantifying Occupational Therapists' Qualitative Evaluations of Postural Control. *Occupational therapy international*. Vol. 2020.

HAMZA, M.; LAROCQUE, D., 2005. An empirical comparison of ensemble methods based on classification trees. *Journal of Statistical Computation and Simulation*. Vol. 75, no. 8, pp. 629–643.

HAN, J. J.; KURILLO, G.; ABRESCH, R. T.; BIE, E. de; NICORICI, A.; BAJCSY, R., 2015. Reachable workspace in facioscapulohumeral muscular dystrophy (FSHD) by Kinect. *Muscle & nerve*. Vol. 51, no. 2, pp. 168–175.

HAN, J.; SHAO, L.; XU, D.; SHOTTON, J., 2013. Enhanced computer vision with microsoft kinect sensor: A review. *IEEE transactions on cybernetics*. Vol. 43, no. 5, pp. 1318–1334.

HARIHARAN, B.; ARBELÁEZ, P.; GIRSHICK, R.; MALIK, J., 2014. Simultaneous detection and segmentation. In: *European Conference on Computer Vision*, pp. 297–312.

HE, K.; GKIOXARI, G.; DOLLÁR, P.; GIRSHICK, R., 2017. Mask r-cnn. In: *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969.

HE, K.; ZHANG, X.; REN, S.; SUN, J., 2016. Deep residual learning for image recognition. In: *in Proc. of the CVPR*, pp. 770–778.

HEISELE, B.; WOEHLER, C., 1998. Motion-based recognition of pedestrians. In: *Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No. 98EX170)*. Vol. 2, pp. 1325–1330.

HERATH, S.; HARANDI, M.; PORIKLI, F., 2017. Going deeper into action recognition: A survey. *Image and vision computing*. Vol. 60, pp. 4–21.

HOCHREITER, S., 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*. Vol. 6, no. 02, pp. 107–116.

HOCHREITER, S.; SCHMIDHUBER, J., 1997. Long short-term memory. *Neural computation*. Vol. 9, no. 8, pp. 1735–1780.

HOLT, G. A. ten; REINDERS, M. J.; HENDRIKS, E., 2007. Multi-dimensional dynamic time warping for gesture recognition. In: *Thirteenth annual conference of the Advanced School for Computing and Imaging*. Vol. 300, p. 1.

HOWARD, A. G. et al., 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *CoRR*. Vol. abs/1704.04861. Available from arXiv: `1704.04861`.

HSIAO, C.-P.; ZHAO, C.; DO, E. Y.-L., 2013. The Digital Box and Block Test Automating traditional post-stroke rehabilitation assessment. In: *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on*, pp. 360–363.

HUANG, G.; LIU, Z.; WEINBERGER, K. Q.; MAATEN, L. van der, 2017. Densely connected convolutional networks. In: *IEEE CVPR*. Vol. 1, pp. 4700–4708.

HUANG, J.-D., 2011. Kinerehab: a kinect-based system for physical rehabilitation: a pilot study for young adults with motor disabilities. In: *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, pp. 319–320.

HUSSEIN, N.; GAVVES, E.; SMEULDERS, A. W., 2017. Unified embedding and metric learning for zero-exemplar event detection. In: *in Proc. of the CVPR*, pp. 1096–1105.

IANDOLA, F. N.; HAN, S.; MOSKEWICZ, M. W.; ASHRAF, K.; DALLY, W. J.; KEUTZER, K., 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size. *arXiv preprint arXiv:1602.07360*.

IM, D. J.; MA, H.; TAYLOR, G.; BRANSON, K., 2018. Quantitatively evaluating GANs with divergences proposed for training. *arXiv preprint arXiv:1803.01045*.

IOFFE, S.; SZEGEDY, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

JADERBERG, M.; SIMONYAN, K.; ZISSERMAN, A., et al., 2015. Spatial transformer networks. In: *in Proc. of the NIPS*, pp. 2017–2025.

JALAL, A.; KAMAL, S.; KIM, D., 2017. *International Journal of Interactive Multimedia and Artificial Intelligence*. Vol. 4, A Depth Video-based Human Detection and Activity Recognition using Multi-features and Embedded Hidden Markov Models for Health Care Monitoring Systems. ImaI-Software. Available also from: `https://www.ijimai.org/journal/node/1516`.

JAMIL, N.; SEMBOK, T. M. T.; BAKAR, Z. A., 2008. Noise removal and enhancement of binary images using morphological operations. In: *2008 International Symposium on Information Technology*. Vol. 4, pp. 1–6.

JÉGOU, H.; DOUZE, M.; SCHMID, C.; PÉREZ, P., 2010. Aggregating local descriptors into a compact image representation. In: *in Proc. of the CVPR*, pp. 3304–3311.

JI, S.; XU, W.; YANG, M.; YU, K., 2012. 3D convolutional neural networks for human action recognition. *IEEE transactions on PAMI*. Vol. 35, no. 1, pp. 221–231.

JI, Y.; YE, G.; CHENG, H., 2014. Interactive body part contrast mining for human interaction recognition. In: *in Proc. of the ICMEW*, pp. 1–6.

JIA DENG; WEI DONG; SOCHER, R.; LI-JIA LI; KAI LI; LI FEI-FEI, 2009. ImageNet: A large-scale hierarchical image database. In: *CVPR*. IEEE, pp. 248–255. ISBN 978-1-4244-3992-8. Available from DOI: `10.1109/CVPR.2009.5206848`.

JUN, S.-k.; KUMAR, S.; ZHOU, X.; RAMSEY, D. K.; KROVI, V. N., 2013. Automation for individualization of Kinect-based quantitative progressive exercise regimen. In: *Automation Science and Engineering (CASE), 2013 IEEE International Conference on*, pp. 243–248.

KARGAR, B. A. H.; MOLLAHOSSEINI, A.; STRUEMPH, T.; PACE, W.; NIELSEN, R. D.; MAHOOR, M. H., 2014. Automatic measurement of physical mobility in get-up-and-go test using Kinect sensor. In: *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3492–3495.

KE, L.; CHANG, M.-C.; QI, H.; LYU, S., 2018. Multi-Scale Structure-Aware Network for Human Pose Estimation. *arXiv preprint arXiv:1803.09894*.

KE, Q.; BENNAMOUN, M.; AN, S.; SOHEL, F.; BOUSSAID, F., 2017. A new representation of skeleton sequences for 3d action recognition. In: *In Proc. of the CVPR*, pp. 3288–3297.

KE, S.-R.; THUC, H. L. U.; LEE, Y.-J.; HWANG, J.-N.; YOO, J.-H.; CHOI, K.-H., 2013. A review on video-based human activity recognition. *Computers*. Vol. 2, no. 2, pp. 88–131.

KEOGH, E.; CHAKRABARTI, K.; PAZZANI, M.; MEHROTRA, S., 2001. Dimensionality reduction for fast similarity search in large time series databases. *Knowledge and information Systems*. Vol. 3, no. 3, pp. 263–286.

KERTÉSZ, C., 2013. Physiotherapy Exercises Recognition Based on RGB-D Human Skeleton Models. In: *Modelling Symposium (EMS), 2013 European*, pp. 21–29.

KHAN, M. H.; HELSPER, J.; FARID, M. S.; GRZEGORZEK, M., 2018. A computer vision-based system for monitoring Vojta therapy. *International journal of medical informatics*. Vol. 113, pp. 85–95.

KHAN, T.; NYHOLM, D.; WESTIN, J.; DOUGHERTY, M., 2014. A computer vision framework for finger-tapping evaluation in Parkinson's disease. *Artificial intelligence in medicine*. Vol. 60, no. 1, pp. 27–40.

KHOSLA, A.; JAYADEVAPRAKASH, N.; YAO, B.; FEI-FEI, L., 2011. Novel Dataset for Fine-Grained Image Categorization. In: *CVPR workshop*. Colorado Springs, CO.

KIM, T. S.; REITER, A., 2017. Interpretable 3d human action analysis with temporal convolutional networks. In: *in Proc. of the CVPRW*, pp. 1623–1631.

KINGMA, D. P.; BA, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

KIPF, T. N.; WELLING, M., 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

KISHORE KUMAR, N.; SCHNEIDER, J., 2017. Literature survey on low rank approximation of matrices. *Linear and Multilinear Algebra*. Vol. 65, no. 11, pp. 2212–2244.

KOCABAS, M.; ATHANASIOU, N.; BLACK, M. J., 2020. VIBE: Video Inference for Human Body Pose and Shape Estimation. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

KOPPULA, H. S.; GUPTA, R.; SAXENA, A., 2013. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*. Vol. 32, no. 8, pp. 951–970.

KOVASHKA, A.; GRAUMAN, K., 2010. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In: *CVPR*. IEEE, pp. 2046–2053. ISBN 9781424469840. ISSN 10636919. Available from DOI: 10.1109/CVPR.2010.5539881.

KRAPAC, J.; VERBEEK, J.; JURIE, F., 2011. Modeling spatial layout with fisher vectors for image categorization. In: *2011 International Conference on Computer Vision*, pp. 1487–1494.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E., 2012. Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*, pp. 1097–1105.

KURILLO, G. et al., 2011. Real-time 3D avatars for tele-rehabilitation in virtual reality. *Medicine Meets Virtual Reality 18: NextMed*. Vol. 163, p. 290.

KURILLO, G.; CHEN, A.; BAJCSY, R.; HAN, J. J., 2013. Evaluation of upper extremity reachable workspace using Kinect camera. *Technology and Health Care*. Vol. 21, no. 6, pp. 641–656.

LAN, Z.; LIN, M.; LI, X.; HAUPTMANN, A. G.; RAJ, B., 2015. Beyond Gaussian Pyramid: Multi-skip Feature Stacking for action recognition. In: *CVPR*. IEEE. Vol. 07-12-June, pp. 204–212. ISBN 9781467369640. ISSN 10636919. Available from DOI: 10.1109/CVPR.2015.7298616.

LAWRENCE, N. D., 2004. Gaussian process latent variable models for visualisation of high dimensional data. In: *Advances in neural information processing systems*, pp. 329–336.

LEA, C.; FLYNN, M. D.; VIDAL, R.; REITER, A.; HAGER, G. D., 2017. Temporal convolutional networks for action segmentation and detection. In: *in Proc. of the CVPR*, pp. 156–165.

LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P., 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*. Vol. 86, no. 11, pp. 2278–2324.

LEE, K.-H., 2015. The role of compensatory movements patterns in spontaneous recovery after stroke. *Journal of physical therapy science*. Vol. 27, no. 9, pp. 2671–2673.

LEIGHTLEY, D.; DARBY, J.; LI, B.; MCPHEE, J. S.; YAP, M. H., 2013. Human activity recognition for physical rehabilitation. In: *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*, pp. 261–266.

LEIGHTLEY, D.; MCPHEE, J. S.; YAP, M. H., 2017a. Automated analysis and quantification of human mobility using a depth sensor. *IEEE journal of biomedical and health informatics*. Vol. 21, no. 4, pp. 939–948.

LEIGHTLEY, D.; MUKHOPADHYAY, S. C.; GHAYVAT, H.; YAP, M. H., 2017b. Deep convolutional neural networks for motion instability identification using kinect. In: *Machine Vision Applications (MVA), 2017 Fifteenth IAPR International Conference on*, pp. 310–313.

LEIGHTLEY, D.; YAP, M. H.; COULSON, J.; BARNOUIN, Y.; MCPHEE, J. S., 2015. Benchmarking human motion analysis using kinect one: An open source dataset. In: *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific*, pp. 1–7.

LEU, A.; RISTIĆ-DURRANT, D.; GRÄSER, A., 2011. A robust markerless vision-based human gait analysis system. In: *Applied Computational Intelligence and Informatics (SACI), 2011 6th IEEE International Symposium on*, pp. 415–420.

LI, B.; MENG, Q.; HOLSTEIN, H., 2004. Articulated pose identification with sparse point features. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*. Vol. 34, no. 3, pp. 1412–1422.

LI, B.; MENG, Q.; HOLSTEIN, H., 2008. Articulated motion reconstruction from feature points. *Pattern Recognition*. Vol. 41, no. 1, pp. 418–431.

LI, B.; DAI, Y.; CHENG, X.; CHEN, H.; LIN, Y.; HE, M., 2017a. Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN. In: *in Proc. of the ICMEW*, pp. 601–604.

LI, C.; WANG, P.; WANG, S.; HOU, Y.; LI, W., 2017b. Skeleton-based action recognition using LSTM and CNN. In: *2017 IEEE International Conference on Multimedia and Expo Workshops, ICMEW 2017*. IEEE, pp. 585–590. ISBN 9781538605608. Available from DOI: 10.1109/ICMEW.2017.8026287.

LI, J.; WANG, C.; ZHU, H.; MAO, Y.; FANG, H.-S.; LU, C., 2018a. CrowdPose: Efficient Crowded Scenes Pose Estimation and A New Benchmark. *arXiv preprint arXiv:1812.00324*.

LI, L.; VAKANSKI, A., 2018. Generative adversarial networks for generation and classification of physical rehabilitation movement episodes. *International journal of machine learning and computing*. Vol. 8, no. 5, p. 428.

LI, M. H.; MESTRE, T. A.; FOX, S. H.; TAATI, B., 2018b. Vision-based assessment of parkinsonism and levodopa-induced dyskinesia with pose estimation. *Journal of neuroengineering and rehabilitation*. Vol. 15, no. 1, p. 97.

LI, W.; ZHANG, Z.; LIU, Z., 2010. Action recognition based on a bag of 3d points. In: *CVPR Workshops (CVPRW)*, pp. 9–14.

LI, Y.; JI, B.; SHI, X.; ZHANG, J.; KANG, B.; WANG, L., 2020. Tea: Temporal excitation and aggregation for action recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 909–918.

LIAO, Y.; VAKANSKI, A.; XIAN, M., 2019. A deep learning framework for assessment of quality of rehabilitation exercises. *arXiv preprint arXiv:1901.10435*.

LIN, L.; WANG, K.; ZUO, W.; WANG, M.; LUO, J.; ZHANG, L., 2016. A deep structured model with radius–margin bound for 3D human activity recognition. *IJCV*. Vol. 118, no. 2, pp. 256–273.

LIN, M.; CHEN, Q.; YAN, S., 2013a. Network in network. *arXiv preprint arXiv:1312.4400*.

LIN, T. Y.; HSIEH, C. H.; LEE, J. D., 2013b. A kinect-based system for physical rehabilitation: Utilizing Tai Chi exercises to improve movement disorders in patients with balance ability. In: *Proceedings - Asia Modelling Symposium 2013: 7th Asia International Conference on Mathemat-*

*ical Modelling and Computer Simulation, AMS 2013*. IEEE, pp. 149–153. ISBN 9780769551012. Available from DOI: 10.1109/AMS.2013.29.

LIN, T.-Y.; HSIEH, C.-H.; LEE, J.-D., 2013c. A kinect-based system for physical rehabilitation: Utilizing tai chi exercises to improve movement disorders in patients with balance ability. In: *Modelling Symposium (AMS), 2013 7th Asia*, pp. 149–153.

LIU, D.; BELLOTTO, N.; YUE, S., 2019a. Deep Spiking Neural Network for Video-Based Disguise Face Recognition Based on Dynamic Facial Movements. *IEEE Transactions on Neural Networks and Learning Systems*.

LIU, J.; SHAHROUDY, A.; PEREZ, M.; WANG, G.; DUAN, L.-Y.; KOT, A. C., 2019b. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on PAMI*. Vol. 42, no. 10, pp. 2684–2701.

LIU, J.; SHAHROUDY, A.; XU, D.; WANG, G., 2016a. Spatio-temporal LSTM with trust gates for 3D human action recognition. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. ECCV 2016. Vol. 9907 LNCS, pp. 816–833. ISBN 9783319464862. ISSN 16113349. Available from DOI: 10.1007/978-3-319-46487-9_50.

LIU, J.; SHAHROUDY, A.; XU, D.; WANG, G., 2016b. Spatio-temporal lstm with trust gates for 3d human action recognition. In: *in Proc. of the AAAI*, pp. 816–833.

LIU, M.; LIU, H.; CHEN, C., 2017. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*. Vol. 68, pp. 346–362.

LIU, T. T.; HSIEH, C. T.; CHUNG, R. C.; WANG, Y. S., 2013. Physical rehabilitation assistant system based on Kinect. In: *Applied Mechanics and Materials*. Vol. 284, pp. 1686–1690.

LIU, Z.; SARKAR, S., 2005. Effect of silhouette quality on hard problems in gait recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*. Vol. 35, no. 2, pp. 170–183.

LOWE, D. G., 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*. Vol. 60, no. 2, pp. 91–110.

LUBLINERMAN, R.; ÖZAY, N.; ZARPALAS, D.; CAMPS, O., 2006. Activity recognition from silhouettes using linear systems and model (In)validation techniques. In: *Proceedings - International Conference on Pattern Recognition*. IEEE. Vol. 1, pp. 347–350. ISBN 0769525210. ISSN 10514651. Available from DOI: 10.1109/ICPR.2006.210.

LUCAS, B. D.; KANADE, T., 1981. An Iterative Image Registration Technique with an Application to Stereo Vision, pp. 674–679. Available also from: http://dl.acm.org/citation.cfm?id=1623264.1623280.

LUO, J.; WANG, W.; QI, H., 2013. Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In: *in Proc. of the ICCV*, pp. 1809–1816.

LUONG, M.-T.; PHAM, H.; MANNING, C. D., 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

LUVIZON, D. C.; TABIA, H.; PICARD, D., 2016. Learning features combination for human action recognition from skeleton sequences. *Pattern Recognition Letters*. Vol. 99, pp. 13–20. ISSN 01678655. Available from DOI: 10.1016/j.patrec.2017.02.001.

MA, M.; FAN, H.; KITANI, K. M., 2016. Going deeper into first-person activity recognition. In: *in Proc. of the CVPR*, pp. 1894–1903.

MAATEN, L. v. d.; HINTON, G., 2008. Visualizing data using t-SNE. *Journal of machine learning research*. Vol. 9, no. Nov, pp. 2579–2605.

MAERTENS, G.; SOROUSHIAN, K., 2007. *Accurate and error resilient time stamping method and/or apparatus for the audio-video interleaved (AVI) format*. Google Patents. US Patent App. 11/230,734.

MAKANSI, O.; ILG, E.; CICEK, O.; BROX, T., 2019. Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7144–7153.

MARCO, L.; FARINELLA, G. M., 2018. *Computer Vision for Assistive Healthcare*. Academic Press.

MARTINEZ, J.; HOSSAIN, R.; ROMERO, J.; LITTLE, J. J., 2017. A simple yet effective baseline for 3d human pose estimation. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2640–2649.

MARVIN, K.; ZELTZER, L., 2015. Barthel Index (BI) - Stroke Engine. Available also from: https://www.strokengine.ca/assess/bi/.

MASOUD, O.; PAPANIKOLOPOULOS, N., 2003. A method for human action recognition. *Image and Vision Computing*. Vol. 21, no. 8, pp. 729–743. ISBN 0262-8856. ISSN 02628856. Available from DOI: 10.1016/S0262-8856(03)00068-4.

MATSUO, K.; YAMADA, K.; UENO, S.; NAITO, S., 2014. An attention-based activity recognition for egocentric video. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 551–556.

MELLOULI, D.; HAMDANI, T. M.; AYED, M. B.; ALIMI, A. M., 2017. Morph-CNN: a morphological convolutional neural network for image classification. In: *International Conference on Neural Information Processing*, pp. 110–117.

METCALF, C. D.; ROBINSON, R.; MALPASS, A. J.; BOGLE, T. P.; DELL, T. A.; HARRIS, C.; DEMAIN, S. H., 2013. Markerless motion capture and measurement of hand kinematics: Validation and application to home-based upper limb rehabilitation. *IEEE Transactions on Biomedical Engineering*. Vol. 60, no. 8, pp. 2184–2192. ISBN 1558-2531. ISSN 00189294. Available from DOI: 10.1109/TBME.2013.2250286.

MICROSOFT, 2012a. *Kinect for Windows Runtime 2.0*. Available also from: https://www.microsoft.com/en-us/download/details.aspx?id=44559.

MICROSOFT, 2012b. *Pykinect*. Available also from: https://pypi.org/project/pykinect/.

MIECH, A.; LAPTEV, I.; SIVIC, J., 2017. Learnable pooling with context gating for video classification. *arXiv preprint arXiv:1706.06905*.

MOESLUND Thomas, B.; GRANUM, E., 2001. A Survey of Computer Vision-Based Human Motion Capture. *CVIU*. Vol. 81, no. 3, pp. 231–268. ISSN 1077-3142. Available from DOI: 10.1006/CVIU.2000.0897.

MOLCHANOV, P.; YANG, X.; GUPTA, S.; KIM, K.; TYREE, S.; KAUTZ, J., 2016. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In: *in Proc. of the CVPR*, pp. 4207–4215.

MOUSAVI, H.; KHADEMI, M., 2014. A review on technical and clinical impact of microsoft kinect on physical therapy and rehabilitation. *Journal of medical engineering*. Vol. 2014.

MÜLLER, M., 2007. Dynamic time warping. *Information retrieval for music and motion*, pp. 69–84.

NARASIMHAN, M.; VIOLA, P.; SHILMAN, M., 2006. Online decoding of markov models under latency constraints. In: *Proceedings of the 23rd international conference on Machine learning*, pp. 657–664.

NATARAJAN, S. K.; WANG, X.; SPRANGER, M.; GRÄSER, A., 2017. Reha@ Home-a vision based markerless gait analysis system for rehabilitation at home. In: *Biomedical Engineering (BioMed), 2017 13th IASTED International Conference on*, pp. 32–41.

NCBI, 2020. Hemiplegia/hemiparesis. Available also from: https://www.ncbi.nlm.nih.gov/medgen/852561.

NEGIN, F.; ÖZDEMIR, F.; AKGÜL, C. B.; YÜKSEL, K. A.; ERÇIL, A., 2013. A decision forest based feature selection framework for action recognition from rgb-depth cameras. In: *International conference image analysis and recognition*, pp. 648–657.

NEWELL, A.; YANG, K.; DENG, J., 2016. Stacked hourglass networks for human pose estimation. In: *ECCV*, pp. 483–499.

NG, A. Y.; JORDAN, M. I.; WEISS, Y., 2002. On spectral clustering: Analysis and an algorithm. In: *Advances in neural information processing systems*, pp. 849–856.

NHS, 2017. *Elbow Pain*. Available also from: https://www.nhs.uk/conditions/elbow-and-arm-pain/.

NHS, 2018. *Ataxia*. Available also from: https://www.nhs.uk/conditions/ataxia/.

NIE, B. X.; WEI, P.; ZHU, S.-C., 2017. Monocular 3d human pose estimation by predicting depth on joints. In: *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 3467–3475.

NINDS, 2020. *Tremors*. National Instituite of Neurological Disorders and Stroke. Available also from: https://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/Fact-Sheets/Tremor-Fact-Sheet.

NING, G.; ZHANG, Z.; HE, Z., 2017a. Knowledge-Guided Deep Fractal Neural Networks for Human Pose Estimation. *ICCV*. Available from arXiv: 1705.02407.

NING, G.; ZHANG, Z.; HE, Z., 2017b. Knowledge-Guided Deep Fractal Neural Networks for Human Pose Estimation. *IEEE Transactions on Multimedia*.

*Non Linear Dimensionality Reduction*, 2020. Wikipedia. Available also from: https://en.wikipedia.org/wiki/Nonlinear%5C_dimensionality%5C_reduction.

OBDRŽÁLEK, S.; KURILLO, G.; HAN, J.; ABRESCH, T.; BAJCSY, R., et al., 2012. Real-time human pose detection and tracking for tele-rehabilitation in virtual reality. *Studies in health technology and informatics*. Vol. 173, pp. 320–324.

OLESH, E. V.; YAKOVENKO, S.; GRITSENKO, V., 2014. Automated assessment of upper extremity movement impairment due to stroke. *PloS one*. Vol. 9, no. 8, e104487.

OORD, A. v. d.; DIELEMAN, S.; ZEN, H.; SIMONYAN, K.; VINYALS, O.; GRAVES, A.; KALCHBRENNER, N.; SENIOR, A.; KAVUKCUOGLU, K., 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

OTSU, N., 1979. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*. Vol. 9, no. 1, pp. 62–66.

OTTE, K.; KAYSER, B.; MANSOW-MODEL, S.; VERREL, J.; PAUL, F.; BRANDT, A. U.; SCHMITZ-HÜBSCH, T., 2016. Accuracy and reliability of the kinect version 2 for clinical measurement of motor function. *PloS one*. Vol. 11, no. 11, e0166532.

OYEDOTUN, O. K.; KHASHMAN, A., 2017. Deep learning in vision-based static hand gesture recognition. *Neural Computing and Applications*. Vol. 28, no. 12, pp. 3941–3951.

PAIEMENT, A.; TAO, L., 2014. Online quality assessment of human movement from skeleton data. *BMVA*, pp. 1–12. ISBN 1-901725-52-9. Available from DOI: 10.5244/C.28.79.

PAIEMENT, A.; TAO, L.; HANNUNA, S.; CAMPLANI, M.; DAMEN, D.; MIRMEHDI, M., 2014. Online quality assessment of human movement from skeleton data. In: *BMVA*, pp. 153–166.

PALMA, C.; SALAZAR, A.; VARGAS, F., 2016a. HMM and DTW for Evaluation of therapeutical gestures using kinect. *arXiv preprint arXiv:1602.03742*.

PALMA, C.; SALAZAR, A.; VARGAS, F., 2016b. HMM and DTW for evaluation of therapeutical gestures using kinect. Available from arXiv: 1602.03742.

PAPANDREOU et al., 2017a. Towards Accurate Multi-Person Pose Estimation in the Wild. In: *Proc. CVPR*, pp. 4903–4911.

PAPANDREOU, G.; ZHU, T.; KANAZAWA, N.; TOSHEV, A.; TOMPSON, J.; BREGLER, C.; MURPHY, K., 2017b. Towards Accurate Multi-person Pose Estimation in the Wild. Available from DOI: 10.1109/CVPR.2017.395.

PARRY, I.; CARBULLIDO, C.; KAWADA, J.; BAGLEY, A.; SEN, S.; GREENHALGH, D.; PALMIERI, T., 2014. Keeping up with video game technology: Objective analysis of Xbox Kinect™ and PlayStation 3 Move™ for use in burn rehabilitation. *Burns*. Vol. 40, no. 5, pp. 852–859.

PAVLAKOS, G.; ZHOU, X.; DANIILIDIS, K., 2018. Ordinal depth supervision for 3d human pose estimation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7307–7316.

PAVLLO, D.; FEICHTENHOFER, C.; GRANGIER, D.; AULI, M., 2019. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*.

PEI, W.; XU, G.; LI, M.; DING, H.; ZHANG, S.; LUO, A., 2016. A motion rehabilitation self-training and evaluation system using Kinect. In: *Ubiquitous Robots and Ambient Intelligence (URAI), 2016 13th International Conference on*, pp. 353–357.

PENG, X.; WANG, L.; CAI, Z.; QIAO, Y., 2014. Action and gesture temporal spotting with super vector representation. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. ECCV. Vol. 8925, pp. 518–527. ISBN 9783319161778. ISSN 16113349. Available from DOI: 10.1007/978-3-319-16178-5_36.

PERRONNIN, F.; DANCE, C., 2007. Fisher kernels on visual vocabularies for image categorization. In: *in Proc. of the CVPR*, pp. 1–8.

PINTARIC, T.; KAUFMANN, H., 2007. Affordable infrared-optical pose-tracking for virtual and augmented reality. In: *Proceedings of Trends and Issues in Tracking for Virtual Environments Workshop, IEEE VR*, pp. 44–51.

PIRKER, W.; KATZENSCHLAGER, R., 2017. Gait disorders in adults and the elderly. *Wiener Klinische Wochenschrift*. Vol. 129, no. 3-4, pp. 81–95.

PISHCHULIN, L. et al., 2016. Deepcut: Joint subset partition and labeling for multi person pose estimation. In: *Proc. CPVR 2016*, pp. 4929–4937.

POGRZEBA, L.; NEUMANN, T.; WACKER, M.; JUNG, B., 2018. Analysis and quantification of repetitive motion in long-term rehabilitation. *IEEE journal of biomedical and health informatics*. Vol. 23, no. 3, pp. 1075–1085.

POLANA, R.; NELSON, R., 1994a. Low level recognition of human motion (or how to get your man without finding his body parts). In: *Proceedings of 1994 IEEE Workshop on Motion of Non-rigid and Articulated Objects*, pp. 77–82.

POLANA, R.; NELSON, R., 1994b. Recognizing activities. In: *Proceedings of 12th International Conference on Pattern Recognition*. Vol. 1, pp. 815–818.

RADFORD, A.; METZ, L.; CHINTALA, S., 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

RAHMAN, S.; IRFAN, M.; RAZA, M.; MOYEEZULLAH GHORI, K.; YAQOOB, S.; AWAIS, M., 2020. Performance analysis of boosting classifiers in recognizing activities of daily living. *International Journal of Environmental Research and Public Health*. Vol. 17, no. 3, p. 1082.

RAMAKRISHNA, V.; MUNOZ, D.; HEBERT, M.; BAGNELL, J. A.; SHEIKH, Y., 2014. Pose machines: Articulated pose estimation via inference machines. In: *ECCV*, pp. 33–47.

RATING SCALES FOR PARKINSON'S DISEASE, M. D. S. T. F. on, 2003. The unified Parkinson's disease rating scale (UPDRS): status and recommendations. *Movement Disorders*. Vol. 18, no. 7, pp. 738–750.

REDDY, Y. C. P.; VISWANATH, P.; REDDY, B. E., 2018. Semi-supervised learning: A brief review. *Int. J. Eng. Technol*. Vol. 7, no. 1.8, p. 81.

REDMON, J.; FARHADI, A., 2017. YOLO9000: Better, Faster, Stronger. In: *CVPR 2017*, pp. 6517–6525.

REIMAN, M. P.; MANSKE, R. C., 2011. The assessment of function: How is it measured? A clinical perspective. *Journal of Manual & Manipulative Therapy*. Vol. 19, no. 2, pp. 91–99.

REN, Y.; HU, K.; DAI, X.; PAN, L.; HOI, S. C.; XU, Z., 2019. Semi-supervised deep embedded clustering. *Neurocomputing*. Vol. 325, pp. 121–130.

RICHTER, J.; WIEDE, C.; LEHMANN, L.; HIRTZ, G., 2017a. Motion evaluation by means of joint filtering for assisted physical therapy. In: *Consumer Electronics-Berlin (ICCE-Berlin), 2017 IEEE 7th International Conference on*, pp. 10–14.

RICHTER, J.; WIEDE, C.; SHINDE, B.; HIRTZ, G., 2017b. Motion Error Classification for Assisted Physical Therapy - A Novel Approach using Incremental Dynamic Time Warping and Normalised Hierarchical Skeleton Joint Data. In: *ICPRAM 2017*.

RIVAS, J. J.; ORIHUELA-ESPINA, F.; PALAFOX, L.; BERTHOUZE, N.; CARMEN LARA, M. del; HERNÁNDEZ-FRANCO, J.; SUCAR, E., 2018. Unobtrusive inference of affective states in virtual rehabilitation from upper limb motions: A feasibility study. *IEEE Transactions on Affective Computing*.

RUSSAKOVSKY, O.; DENG, J.; SU, H.; KRAUSE, J.; SATHEESH, S.; MA, S.; HUANG, Z.; KARPATHY, A.; KHOSLA, A.; BERNSTEIN, M., et al., 2015. Imagenet large scale visual recognition challenge. *IJCV*. Vol. 115, no. 3, pp. 211–252.

RYOO, M. S.; AGGARWAL, J. K., 2009. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In: *ICCV*. IEEE, pp. 1593–1600. ISBN 9781424444205. ISSN 1550-5499. Available from DOI: 10.1109/ICCV.2009.5459361.

SABOUR, S.; FROSST, N.; HINTON, G. E., 2017. Dynamic routing between capsules. In: *Advances in neural information processing systems*, pp. 3856–3866.

SÁNCHEZ, J.; PERRONNIN, F.; MENSINK, T.; VERBEEK, J., 2013. Image classification with the fisher vector: Theory and practice. *IJCV*. Vol. 105, no. 3, pp. 222–245.

SANTILLI, G.; LANEVE, G., 2011. Comparing Neural Networks, Invariant Moments and Mathematical Morphology Performances for the Automatic Object Recognition. In: *34th International Symposium on Remote Sensing of Environment 34th*.

SAPP, B.; TASKAR, B., 2013. MODEC: Multimodal Decomposable Models for Human Pose Estimation. In: *In Proc. CVPR*.

SATHYANARAYANA, S.; SATZODA, R. K.; SATHYANARAYANA, S.; THAMBIPILLAI, S., 2018. Vision-based patient monitoring: a comprehensive review of algorithms and technologies. *Journal of Ambient Intelligence and Humanized Computing*. Vol. 9, no. 2, pp. 225–251. ISSN 1868-5145. Available from DOI: 10.1007/s12652-015-0328-1.

SCHAFER, R. W., 2011. What is a Savitzky-Golay filter?[lecture notes]. *IEEE Signal processing magazine*. Vol. 28, no. 4, pp. 111–117.

SCHERER, R.; WAGNER, J.; MOITZI, G.; MÜLLER-PUTZ, G., 2012. Kinect-based detection of self-paced hand movements: enhancing functional brain mapping paradigms. In: *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pp. 4748–4751.

SCHEZ-SOBRINO, S.; MONEKOSSO, D. N.; REMAGNINO, P.; VALLEJO, D.; GLEZ-MORCILLO, C., 2019. Automatic recognition of physical exercises performed by stroke survivors to improve remote rehabilitation. In: *2019 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pp. 1–6.

SCHÖNAUER, C.; PINTARIC, T.; KAUFMANN, H.; JANSEN-KOSTERINK, S.; VOLLENBROEK-HUTTEN, M., 2011. Chronic pain rehabilitation with a serious game using multimodal input. In: *Virtual Rehabilitation (ICVR), 2011 International Conference on*, pp. 1–8.

SEMPENA, S.; MAULIDEVI, N. U.; ARYAN, P. R., 2011. Human action recognition using Dynamic Time Warping. *Electrical Engineering and Informatics (ICEEI), 2011 International Conference on*, pp. 1–5. ISBN 2155-6822 VO -. ISSN 2155-6822. Available from DOI: 10.1109/ICEEI.2011.6021605.

SETO, S.; ZHANG, W.; ZHOU, Y., 2015. Multivariate time series classification using dynamic time warping template selection for human activity recognition. In: *2015 IEEE Symposium Series on Computational Intelligence*, pp. 1399–1406.

SHAHAM, U.; STANTON, K.; LI, H.; NADLER, B.; BASRI, R.; KLUGER, Y., 2018. Spectralnet: Spectral clustering using deep neural networks. *arXiv preprint arXiv:1801.01587*.

SHAHROUDY, A.; LIU, J.; NG, T.-T.; WANG, G., 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: *in Proc. of the CVPR*, pp. 1010–1019.

SHAHROUDY, A.; NG, T.-T.; GONG, Y.; WANG, G., 2017. Deep multimodal feature analysis for action recognition in rgb+ d videos. *IEEE transactions on PAMI*. Vol. 40, no. 5, pp. 1045–1058.

SHAHROUDY, A.; NG, T.-T.; YANG, Q.; WANG, G., 2015. Multimodal multipart learning for action recognition in depth videos. *IEEE transactions on PAMI*. Vol. 38, no. 10, pp. 2123–2129.

SHARMA, S.; KIROS, R.; SALAKHUTDINOV, R., 2016. Action recognition using visual attention. *in Proc. of the ICLRW*.

SHI, J.; MALIK, J., 1998. Motion segmentation and tracking using normalized cuts. In: *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, pp. 1154–1160.

SHI, L.; ZHANG, Y.; CHENG, J.; LU, H., 2019. Skeleton-based action recognition with directed graph neural networks. In: *in Proc. of the CVPR*, pp. 7912–7921.

SHI, Z.; KIM, T.-K., 2017. Learning and Refining of Privileged Information-based RNNs for Action Recognition from Depth Sequences. Available from DOI: 10.1109/CVPR.2017.498.

SHOTTON, J.; FITZGIBBON, A.; COOK, M.; SHARP, T.; FINOCCHIO, M.; MOORE, R.; KIPMAN, A.; BLAKE, A., 2011. Real-time human pose recognition in parts from single depth images. In: *CVPR 2011*, pp. 1297–1304.

SIFRE, L.; MALLAT, S., 2014. Rigid-motion scattering for image classification. *Ph. D. thesis*.

SIGURDSSON, G. A.; VAROL, G.; WANG, X.; FARHADI, A.; LAPTEV, I.; GUPTA, A., 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In: *European Conference on Computer Vision*, pp. 510–526.

SIMONYAN, K.; ZISSERMAN, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR*. Vol. abs/1409.1556.

SIMONYAN, K.; ZISSERMAN, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In: *International Conference on Learning Representations*.

SINGH, B.; MARKS, T. K.; JONES, M.; TUZEL, O.; SHAO, M., 2016. A Multi-stream Bi-directional Recurrent Neural Network for Fine-Grained Action Detection. In: *CVPR*. IEEE, pp. 1961–1970. ISBN 978-1-4673-8851-1. ISSN 10636919. Available from DOI: 10.1109/CVPR.2016.216.

SMITH, D. C.; KORNELSON, K. A., 2013. A comparison of Fisher vectors and Gaussian supervectors for document versus non-document image classification. In: *Applications of Digital Image Processing XXXVI*. Vol. 8856, 88560N.

SONG, S.; LAN, C.; XING, J.; ZENG, W.; LIU, J., 2016. An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data. *Thirty-First AAAI Conference on Artificial Intelligence*. Available from arXiv: 1611.06067.

SONG, S.; LAN, C.; XING, J.; ZENG, W.; LIU, J., 2017. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: *in Proc. of the AAAI*.

SORAN, B.; LOWES, L.; STEELE, K. M., 2016. Evaluation of Infants with Spinal Muscular Atrophy Type-I Using Convolutional Neural Networks. In: *ECCV*, pp. 495–507.

SPASOJEVIĆ, S.; ILIĆ, T. V.; MILANOVIĆ, S.; POTKONJAK, V.; RODIĆ, A.; SANTOS-VICTOR, J., 2017. Combined vision and wearable sensors-based system for movement analysis in rehabilitation. *Methods of information in medicine*. Vol. 56, no. 02, pp. 95–111.

SPASOJEVIĆ, S.; SANTOS-VICTOR, J.; ILIĆ, T.; MILANOVIĆ, S.; POTKONJAK, V.; RODIĆ, A., 2015. A vision-based system for movement analysis in medical applications: the example of Parkinson disease. In: *International Conference on Computer Vision Systems*, pp. 424–434.

STONE, E. E.; SKUBIC, M., 2012. Capturing habitual, in-home gait parameter trends using an inexpensive depth camera. In: *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pp. 5106–5109.

STROKE ASSOCIATION UK, 2017. State of the Nation: Stroke statistics. Available also from: https://www.stroke.org.uk/.

SU, C.-J.; CHIANG, C.-Y.; HUANG, J.-Y., 2014. Kinect-enabled home-based rehabilitation system using Dynamic Time Warping and fuzzy logic. *Applied Soft Computing*. Vol. 22, pp. 652–666.

SUCAR, L. E.; AZCÁRATE, G.; LEDER, R. S.; REINKENSMEYER, D.; HERNÁNDEZ, J.; SANCHEZ, I.; SAUCEDO, P., 2008a. Gesture therapy: A vision-based system for arm rehabilitation after stroke. In: *International Joint Conference on Biomedical Engineering Systems and Technologies*, pp. 531–540.

SUCAR, L. E.; LUIS, R.; LEDER, R.; HERNÁNDEZ, J.; SÁNCHEZ, I., 2010. Gesture therapy: A vision-based system for upper extremity stroke rehabilitation. In: *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, pp. 3690–3693.

SUCAR, L. E.; AZCÁRATE, G.; LEDER, R. S.; REINKENSMEYER, D.; HERNÁNDEZ, J.; SANCHEZ, I.; SAUCEDO, P., 2008b. Gesture therapy: A vision-based system for arm rehabilitation after stroke. In: *Communications in Computer and Information Science*. Springer, Berlin, Heidelberg. Vol. 25 CCIS, pp. 531–540. ISBN 3540922180. ISSN 18650929. Available from DOI: 10.1007/978-3-540-92219-3_40.

SUDHAKARAN, S.; ESCALERA, S.; LANZ, O., 2019. Lsta: Long short-term attention for egocentric action recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9954–9963.

SUMA, E. A.; LANGE, B.; RIZZO, A. S.; KRUM, D. M.; BOLAS, M., 2011. Faast: The flexible action and articulated skeleton toolkit. In: *Virtual Reality Conference (VR), 2011 IEEE*, pp. 247–248.

SUN, X.; WU, P.; HOI, S. C., 2018. Face detection using deep learning: An improved faster RCNN approach. *Neurocomputing*. Vol. 299, pp. 42–50.

SUNG, J.; PONCE, C.; SELMAN, B.; SAXENA, A., 2011. Human activity detection from RGBD images. In: *Workshops at the twenty-fifth AAAI conference on artificial intelligence*.

SZEGEDY, C. et al., 2016. Rethinking the inception architecture for computer vision. In: *In Proc. CVPR*, pp. 2818–2826.

SZEGEDY, C.; IOFFE, S.; VANHOUCKE, V.; ALEMI, A. A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In: *in Proc. of the AAAI*.

SZEGEDY, C.; LIU, W.; JIA, Y.; SERMANET, P.; REED, S.; ANGUELOV, D.; ERHAN, D.; VANHOUCKE, V.; RABINOVICH, A., 2015. Going deeper with convolutions. In: *in Proc. of the CVPR*, pp. 1–9.

TAATI, B.; WANG, R.; HUQ, R.; SNOEK, J.; MIHAILIDIS, A., 2012. Vision-based posture assessment to detect and categorize compensation during robotic rehabilitation therapy. In: *Biomedical Robotics and Biomechatronics (BioRob), 2012 4th IEEE RAS & EMBS International Conference on*, pp. 1607–1613.

TAN, M.; LE, Q. V., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*.

TANG, Y.; TIAN, Y.; LU, J.; LI, P.; ZHOU, J., 2018. Deep progressive reinforcement learning for skeleton-based action recognition. In: *in Proc. of the CVPR*, pp. 5323–5332.

TAO, L.; PAIEMENT, A.; DAMEN, D.; MIRMEHDI, M.; HANNUNA, S.; CAMPLANI, M.; BURGHARDT, T.; CRADDOCK, I., 2016. A comparative study of pose representation and dynamics modelling for online motion quality assessment. *CVIU*. Vol. 148, pp. 136–152.

TAO, L.; VIDAL, R., 2015. Moving poselets: A discriminative and interpretable skeletal motion representation for action recognition. In: *in Proc. of the CVPRW*, pp. 61–69.

TAO, Y.; HU, H., 2004. Colour based human motion tracking for home-based rehabilitation. In: *Systems, Man and Cybernetics, 2004 IEEE International Conference on*. Vol. 1, pp. 773–778.

THOMAS, G.; GADE, R.; MOESLUND, T. B.; CARR, P.; HILTON, A., 2017. Computer vision for sports: Current applications and research topics. *Computer Vision and Image Understanding*. Vol. 159, pp. 3–18.

TITTERINGTON, D. M.; SMITH, A. F.; MAKOV, U. E., 1985. *Statistical analysis of finite mixture distributions*. Wiley,

TOMPSON, J.; GOROSHIN, R.; JAIN, A.; LECUN, Y.; BREGLER, C., 2015. Efficient object localization using convolutional networks. In: *In Proc. CVPR*, pp. 648–656.

TORMENE, P.; GIORGINO, T.; QUAGLINI, S.; STEFANELLI, M., 2009. Matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation. *Artificial intelligence in medicine*. Vol. 45, no. 1, pp. 11–34.

TOSHEV, A.; SZEGEDY, C., 2014. Deeppose: Human pose estimation via deep neural networks. In: *In Proc. CVPR*, pp. 1653–1660.

TRAN, D.; BOURDEV, L.; FERGUS, R.; TORRESANI, L.; PALURI, M., 2015. Learning spatiotemporal features with 3d convolutional networks. In: *in Proc. of the ICCV*, pp. 4489–4497.

TSAI, D.-M.; CHIU, W.-Y.; LEE, M.-H., 2015. Optical flow-motion history image (OF-MHI) for action recognition. *Signal, Image and Video Processing*. Vol. 9, no. 8, pp. 1897–1906.

TU, G.; LI, Q.; JIANG, D., 2019. Dynamic Gesture Recognition Based on HMM-DTW Model Using Leap Motion. In: *International Symposium on Intelligence Computation and Applications*, pp. 788–798.

UDDIN, M. A.; JOOLEE, J. B.; ALAM, A.; LEE, Y.-K., 2017. Human Action Recognition using Adaptive Local Motion Descriptor in Spark. *IEEE Access*. Vol. 5, pp. 1–1. ISSN 2169-3536. Available from DOI: 10.1109/ACCESS.2017.2759225.

ULLAH, A.; MUHAMMAD, K.; DEL SER, J.; BAIK, S. W.; ALBUQUERQUE, V. H. C. de, 2018. Activity recognition using temporal optical flow convolutional features and multilayer LSTM. *IEEE Transactions on Industrial Electronics*. Vol. 66, no. 12, pp. 9692–9702.

VAKANSKI, A.; FERGUSON, J.; LEE, S., 2016. Mathematical modeling and evaluation of human motions in physical therapy using mixture density neural networks. *Journal of physiotherapy & physical rehabilitation*. Vol. 1, no. 4.

VAKANSKI, A.; JUN, H.-p.; PAUL, D.; BAKER, R., 2018. A Data Set of Human Body Movements for Physical Rehabilitation Exercises. *Data*. Vol. 3, no. 1, p. 2.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I., 2017. Attention is all you need. In: *in Proc. of the NIPS*, pp. 5998–6008.

VEMULAPALLI, R.; ARRATE, F.; CHELLAPPA, R., 2014. Human action recognition by representing 3d skeletons as points in a lie group. In: *in Proc. of the CVPR*, pp. 588–595.

VENUGOPALAN, J.; CHENG, C.; STOKES, T. H.; WANG, M. D., 2013. Kinect-based rehabilitation system for patients with traumatic brain injury. In: *Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pp. 4625–4628.

VIOLA, P.; JONES, M. J., 2004. Robust real-time face detection. *International journal of computer vision*. Vol. 57, no. 2, pp. 137–154.

VOULODIMOS, A.; DOULAMIS, N.; DOULAMIS, A.; PROTOPAPADAKIS, E., 2018. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*. Vol. 2018.

VRIGKAS, M.; NIKOU, C.; KAKADIARIS, I. A., 2015. A review of human activity recognition methods. *Frontiers in Robotics and AI*. Vol. 2, p. 28.

WALT, S. v. d.; COLBERT, S. C.; VAROQUAUX, G., 2011. The NumPy array: a structure for efficient numerical computation. *Computing in science & engineering*. Vol. 13, no. 2, pp. 22–30.

WAN, E. A.; VAN DER MERWE, R., 2000. The unscented Kalman filter for nonlinear estimation. In: *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No. 00EX373)*, pp. 153–158.

WANG, J.; LIU, Z.; WU, Y.; YUAN, J., 2012. Mining actionlet ensemble for action recognition with depth cameras. In: *in Proc. of the CVPR*, pp. 1290–1297.

WANG, J.; NIE, X.; XIA, Y.; WU, Y.; ZHU, S.-C., 2014a. Cross-view action modeling, learning and recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2649–2656.

WANG, J.; WU, Y., 2013. Learning maximum margin temporal warping for action recognition. In: *in Proc. of the ICCV*, pp. 2688–2695.

WANG, J.; YANG, Y.; MAO, J.; HUANG, Z.; HUANG, C.; XU, W., 2016. Cnn-rnn: A unified framework for multi-label image classification. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2285–2294.

WANG, J.; YU, L.; WANG, J.; GUO, L.; GU, X.; FANG, Q., 2014b. Automated Fugl-Meyer assessment using SVR model. In: *2014 IEEE International Symposium on Bioelectronics and Bioinformatics (IEEE ISBB 2014)*, pp. 1–4.

WANG, R.; MEDIONI, G.; WINSTEIN, C. J.; BLANCO, C., 2013a. Home monitoring musculoskeletal disorders with a single 3D sensor. In: *CVPR*. IEEE, pp. 521–528. ISBN 9780769549903. ISSN 21607508. Available from DOI: 10.1109/CVPRW.2013.83.

WANG, R.; MEDIONI, G.; WINSTEIN, C.; BLANCO, C., 2013b. Home monitoring musculo-skeletal disorders with a single 3d sensor. In: *CVPR Workshops*, pp. 521–528.

WANG, S. B.; QUATTONI, A.; MORENCY, L.-P.; DEMIRDJIAN, D.; DARRELL, T., 2006a. Hidden conditional random fields for gesture recognition. In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. Vol. 2, pp. 1521–1527.

WANG, T.; CHO, K.; WEN, M., 2019. Attention-based mixture density recurrent networks for history-based recommendation. In: *Proceedings of the 1st International Workshop on Deep Learning Practice for High-Dimensional Sparse Data*, pp. 1–9.

WANG, Y.; JIANG, H.; DREW, M. S.; LI, Z. N.; MORI, G., 2006b. Unsupervised discovery of action classes. In: *CVPR*. IEEE. Vol. 2, pp. 1654–1661. ISBN 0769525970. ISSN 10636919. Available from DOI: 10.1109/CVPR.2006.321.

WEBSTER, D.; CELIK, O., 2014. Systematic review of Kinect applications in elderly care and stroke rehabilitation. *Journal of neuroengineering and rehabilitation*. Vol. 11, no. 1, p. 108.

WEI, S.-E.; RAMAKRISHNA, V.; KANADE, T.; SHEIKH, Y., 2016. Convolutional pose machines. In: *Proc. CVPR*, pp. 4724–4732.

WEI, Y.; LI, W.; FAN, Y.; XU, L.; CHANG, M.-C.; LYU, S., 2020. 3D Single-Person Concurrent Activity Detection Using Stacked Relation Network. *AAAI*.

WELCH, P., 1967. The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on audio and electroacoustics*. Vol. 15, no. 2, pp. 70–73.

WILLIAMS, C.; VAKANSKI, A.; LEE, S.; PAUL, D., 2019. Assessment of physical rehabilitation movements through dimensionality reduction and statistical modeling. *Medical engineering & physics*. Vol. 74, pp. 13–22.

WU, D.; PIGOU, L.; KINDERMANS, P.-J.; LE, N. D.-H.; SHAO, L.; DAMBRE, J.; ODOBEZ, J.-M., 2016a. Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE transactions on PAMI*. Vol. 38, no. 8, pp. 1583–1597.

WU, H.; PAN, W.; XIONG, X.; XU, S., 2014. Human activity recognition based on the combined SVM&HMM. In: *Information and Automation (ICIA), 2014 IEEE International Conference on*. IEEE, pp. 219–224. ISBN 978-1-4799-4100-1. Available from DOI: 10.1109/ICInfA.2014.6932656.

WU, J.; XUE, T.; LIM, J. J.; TIAN, Y.; TENENBAUM, J. B.; TORRALBA, A.; FREEMAN, W. T., 2016b. Single image 3d interpreter network. In: *European Conference on Computer Vision*, pp. 365–382.

XIA, L.; CHEN, C. C.; AGGARWAL, J. K., 2012a. View invariant human action recognition using histograms of 3D joints. In: *CVPR Workshops*. IEEE, pp. 20–27. ISBN 9781467316118. ISSN 21607508. Available from DOI: 10.1109/CVPRW.2012.6239233.

XIA, L.; CHEN, C.-C.; AGGARWAL, J. K., 2012b. View invariant human action recognition using histograms of 3d joints. In: *CVPR Workshops*, pp. 20–27.

XIAO, X.; WAN, W., 2017. Human pose estimation via improved ResNet50. In: *ICSSC 2017*, pp. 1–5. Available from DOI: 10.1049/cp.2017.0126.

XING, Y.; LV, C.; WANG, H.; CAO, D.; VELENIS, E.; WANG, F.-Y., 2019. Driver activity recognition for intelligent vehicles: A deep learning approach. *IEEE Transactions on Vehicular Technology*. Vol. 68, no. 6, pp. 5379–5390.

XU, K.; BA, J.; KIROS, R.; CHO, K.; COURVILLE, A.; SALAKHUDINOV, R.; ZEMEL, R.; BENGIO, Y., 2015a. Show, attend and tell: Neural image caption generation with visual attention. In: *in Proc. of the ICML*, pp. 2048–2057.

XU, Y.; CHENG, J.; WANG, L.; XIA, H.; LIU, F.; TAO, D., 2018. Ensemble one-dimensional convolution neural networks for skeleton-based action recognition. *IEEE Signal Processing Letters*. Vol. 25, no. 7, pp. 1044–1048.

XU, Z.; YANG, Y.; HAUPTMANN, A. G., 2015b. A discriminative CNN video representation for event detection. In: *in Proc. of the CVPR*, pp. 1798–1807.

YAMATO, J.; OHYA, J.; ISHII, K., 1992. Recognizing human action in time-sequential images using hidden Markov model. In: *CVPR*. IEEE Comput. Soc. Press, pp. 379–385. ISBN 0-8186-2855-3. Available from DOI: 10.1109/CVPR.1992.223161.

YAN, S.; XIONG, Y.; LIN, D., 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In: *in Proc. of the AAAI*.

YANG, L.; ZHANG, L.; DONG, H.; ALELAIWI, A.; EL SADDIK, A., 2015a. Evaluating and improving the depth accuracy of Kinect for Windows v2. *IEEE Sensors Journal*. Vol. 15, no. 8, pp. 4275–4285.

YANG, S.; LUO, P.; LOY, C.-C.; TANG, X., 2015b. From facial parts responses to face detection: A deep learning approach. In: *Proceedings of the IEEE international conference on computer vision*, pp. 3676–3684.

YANG, W.; LI, S.; OUYANG, W.; LI, H.; WANG, X., 2017. Learning feature pyramids for human pose estimation. In: *ICCV*. Vol. 2.

YANG, W.; OUYANG, W.; WANG, X.; REN, J.; LI, H.; WANG, X., 2018. 3d human pose estimation in the wild by adversarial learning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5255–5264.

YANG, Y.; RAMANAN, D., 2011. Articulated pose estimation with flexible mixtures-of-parts. In: *CVPR 2011*, pp. 1385–1392. ISSN 1063-6919. Available from DOI: 10.1109/CVPR.2011.5995741.

YE, Q.; KIM, T.-K., 2018. Occlusion-aware hand pose estimation using hierarchical mixture density network. In: *ECCV*, pp. 801–817.

YEUNG, S.; RUSSAKOVSKY, O.; JIN, N.; ANDRILUKA, M.; MORI, G.; FEI-FEI, L., 2018. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision*. Vol. 126, no. 2-4, pp. 375–389.

YOSINSKI, J.; CLUNE, J.; BENGIO, Y.; LIPSON, H., 2014. How transferable are features in deep neural networks? In: *NIPS*, pp. 3320–3328.

YU, X.; XIONG, S., 2019. A dynamic time warping based algorithm to evaluate kinect-enabled home-based physical rehabilitation exercises for older people. *Sensors*. Vol. 19, no. 13, p. 2882.

YUN, K.; HONORIO, J.; CHATTOPADHYAY, D.; BERG, T. L.; SAMARAS, D., 2012. Two-person interaction detection using body-pose features and multiple instance learning. In: *in Proc. of the CVPRW*, pp. 28–35.

ZANFIR, M.; LEORDEANU, M.; SMINCHISESCU, C., 2013. The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In: *in Proc. of the ICCV*, pp. 2752–2759.

ZARIFFA, J.; STEEVES, J. D., 2011. Computer vision-based classification of hand grip variations in neurorehabilitation. In: *Rehabilitation Robotics (ICORR), 2011 IEEE International Conference on*, pp. 1–4.

ZHANG, H.; GOODFELLOW, I.; METAXAS, D.; ODENA, A., 2018. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*.

ZHANG, K.; ZHANG, Z.; LI, Z.; QIAO, Y., 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*. Vol. 23, no. 10, pp. 1499–1503.

ZHANG, L.; ZHU, G.; SHEN, P.; SONG, J.; AFAQ SHAH, S.; BENNAMOUN, M., 2017a. Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 3120–3128.

ZHANG, P.; LAN, C.; XING, J.; ZENG, W.; XUE, J.; ZHENG, N., 2017b. View adaptive recurrent neural networks for high performance human action recognition from skeleton data. In: *in Proc. of the CVPR*, pp. 2117–2126.

ZHANG, Z.; HUANG, K.; TAN, T., 2006. Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes. In: *18th International Conference on Pattern Recognition (ICPR'06)*. Vol. 3, pp. 1135–1138.

ZHENG, F.; WEBB, G. I., 2005. A comparative study of Semi-naive Bayes methods in classification learning. In.

ZHI, Y. X.; LUKASIK, M.; LI, M. H.; DOLATABADI, E.; WANG, R. H.; TAATI, B., 2018. Automatic Detection of Compensation During Robotic Stroke Rehabilitation Therapy. *IEEE journal of translational engineering in health and medicine*. Vol. 6, pp. 1–7.

ZHOU, H.; HU, H., 2008. Human motion tracking for rehabilitation—A survey. *Biomedical Signal Processing and Control*. Vol. 3, no. 1, pp. 1–18.

ZHU, W.; LAN, C.; XING, J.; ZENG, W.; LI, Y.; SHEN, L.; XIE, X., 2016. Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks. In: *in Proc. of the AAAI*.

ZHU, Y.; CHEN, W.; GUO, G., 2015. Fusing multiple features for depth-based action recognition. *ACM Transactions on Intelligent Systems and Technology (TIST)*. Vol. 6, no. 2, pp. 1–20.

ZOPH, B.; VASUDEVAN, V.; SHLENS, J.; LE, Q. V., 2018. Learning transferable architectures for scalable image recognition. In: *in Proc. of the CVPR*, pp. 8697–8710.

# Appendices

# Appendix A

# Additional Data

## A.1 Chapter 5: Multi-label Activity Recognition Dataset

| Activities | Impairments | S001 | S002 | S003 | S004 | S005 | S006 | S007 | S008 | S009 | S010 | Sum |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Answering Phone | Normal | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 10 | 11 | 11 | 109 |
| Answering Phone | Ataxic | 11 | 12 | 11 | 12 | 12 | 12 | 12 | 12 | 11 | 12 | 117 |
| Answering Phone | Elbow Rigidity | 10 | 12 | 11 | 11 | 12 | 12 | 12 | 12 | 10 | 13 | 115 |
| Answering Phone | Tremors | 11 | 11 | 11 | 11 | 11 | 10 | 12 | 11 | 12 | 11 | 111 |
| Answering Phone | Weak Shoulder | 11 | 11 | 11 | 10 | 12 | 12 | 11 | 11 | 14 | 11 | 114 |
| Brushing Floor | Normal | 11 | 12 | 14 | 11 | 11 | 11 | 12 | 11 | 11 | 11 | 115 |
| Brushing Floor | Ataxic | 11 | 12 | 12 | 13 | 11 | 12 | 12 | 12 | 12 | 13 | 120 |
| Brushing Floor | Elbow Rigidity | 11 | 10 | 13 | 14 | 12 | 12 | 12 | 11 | 11 | 12 | 118 |
| Brushing Floor | Tremors | 11 | 11 | 12 | 12 | 12 | 10 | 10 | 11 | 12 | 16 | 117 |
| Brushing Floor | Weak Shoulder | 11 | 10 | 11 | 10 | 11 | 10 | 10 | 11 | 10 | 12 | 106 |
| Brushing Hair | Normal | 12 | 11 | 11 | 11 | 11 | 11 | 10 | 11 | 12 | 11 | 111 |
| Brushing Hair | Ataxic | 11 | 11 | 12 | 12 | 12 | 11 | 12 | 11 | 13 | 12 | 117 |
| Brushing Hair | Elbow Rigidity | 12 | 11 | 10 | 10 | 14 | 11 | 12 | 12 | 12 | 12 | 116 |
| Brushing Hair | Tremors | 11 | 10 | 11 | 10 | 11 | 12 | 12 | 11 | 12 | 12 | 112 |
| Brushing Hair | Weak Shoulder | 12 | 11 | 13 | 10 | 11 | 11 | 11 | 12 | 12 | 12 | 115 |
| Clapping | Normal | 11 | 11 | 10 | 12 | 12 | 11 | 11 | 11 | 13 | 13 | 115 |
| Clapping | Ataxic | 11 | 11 | 11 | 10 | 13 | 12 | 12 | 12 | 12 | 12 | 116 |
| Clapping | Elbow Rigidity | 11 | 12 | 10 | 11 | 12 | 11 | 11 | 11 | 11 | 12 | 112 |
| Clapping | Tremors | 11 | 11 | 11 | 12 | 12 | 12 | 11 | 11 | 12 | 15 | 118 |
| Clapping | Weak Shoulder | 12 | 11 | 11 | 10 | 12 | 10 | 12 | 11 | 13 | 14 | 116 |
| Drinking | Normal | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 12 | 111 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Drinking | Ataxic | 11 | 12 | 11 | 12 | 12 | 12 | 12 | 11 | 12 | 11 | 116 |
| Drinking | Elbow Rigidity | 11 | 10 | 11 | 12 | 12 | 12 | 12 | 12 | 12 | 12 | 116 |
| Drinking | Tremors | 11 | 11 | 11 | 11 | 14 | 11 | 12 | 12 | 12 | 13 | 118 |
| Drinking | Weak Shoulder | 12 | 11 | 12 | 12 | 12 | 12 | 12 | 13 | 10 | 12 | 118 |
| Reaching Above | Normal | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 12 | 12 | 12 | 113 |
| Reaching Above | Ataxic | 12 | 11 | 10 | 13 | 0 | 12 | 11 | 12 | 12 | 13 | 106 |
| Reaching Above | Elbow Rigidity | 12 | 12 | 11 | 10 | 12 | 14 | 12 | 12 | 12 | 12 | 119 |
| Reaching Above | Tremors | 11 | 12 | 11 | 12 | 12 | 11 | 12 | 12 | 11 | 13 | 117 |
| Reaching Above | Weak Shoulder | 12 | 11 | 11 | 11 | 0 | 12 | 12 | 11 | 12 | 12 | 104 |
| Sitting | Normal | 11 | 11 | 11 | 11 | 11 | 12 | 10 | 11 | 12 | 11 | 111 |
| Sitting | Ataxic | 11 | 10 | 11 | 11 | 12 | 12 | 12 | 12 | 13 | 12 | 116 |
| Sitting | Knee Rigidity | 10 | 10 | 9 | 10 | 12 | 11 | 12 | 9 | 14 | 14 | 111 |
| Sitting | Weakness to One Side | 11 | 11 | 12 | 12 | 12 | 11 | 10 | 12 | 14 | 11 | 116 |
| Sitting | Wider Gait | 11 | 11 | 12 | 11 | 11 | 11 | 11 | 11 | 12 | 12 | 113 |
| Standing | Normal | 11 | 11 | 11 | 10 | 11 | 11 | 10 | 10 | 10 | 12 | 107 |
| Standing | Ataxic | 11 | 11 | 11 | 12 | 13 | 12 | 11 | 12 | 11 | 11 | 115 |
| Standing | Knee Rigidity | 10 | 10 | 10 | 0 | 10 | 12 | 12 | 13 | 10 | 13 | 100 |
| Standing | Weakness to One Side | 11 | 14 | 11 | 11 | 12 | 11 | 11 | 11 | 9 | 12 | 113 |
| Standing | Wider Gait | 11 | 12 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 12 | 112 |
| Walking | Normal | 10 | 11 | 11 | 11 | 11 | 11 | 10 | 12 | 12 | 15 | 114 |
| Walking | Ataxic | 11 | 12 | 10 | 11 | 12 | 11 | 13 | 12 | 11 | 15 | 118 |
| Walking | Knee Rigidity | 9 | 10 | 9 | 10 | 11 | 15 | 12 | 10 | 12 | 15 | 113 |
| Walking | Weakness to One Side | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 13 | 11 | 13 | 114 |
| Walking | Wider Gait | 11 | 11 | 11 | 11 | 10 | 10 | 11 | 13 | 12 | 13 | 113 |
| Wearing Glasses | Normal | 11 | 11 | 11 | 10 | 10 | 11 | 10 | 11 | 10 | 11 | 106 |
| Wearing Glasses | Ataxic | 12 | 10 | 12 | 12 | 11 | 12 | 11 | 12 | 12 | 12 | 116 |
| Wearing Glasses | Elbow Rigidity | 11 | 10 | 11 | 14 | 12 | 12 | 12 | 12 | 11 | 11 | 116 |
| Wearing Glasses | Tremors | 11 | 12 | 11 | 12 | 11 | 11 | 11 | 11 | 12 | 11 | 113 |
| Wearing Glasses | Weak Shoulder | 12 | 12 | 13 | 12 | 11 | 11 | 12 | 13 | 11 | 13 | 120 |
| Sum | | 554 | 555 | 556 | 551 | 556 | 571 | 568 | 574 | 582 | 618 | 5685 |

Table A.1: Subject-wise Activity and Imparirment distribution of the number of sequence filmed

## A.2 Chapter 8: Multi Label Activity Recognition

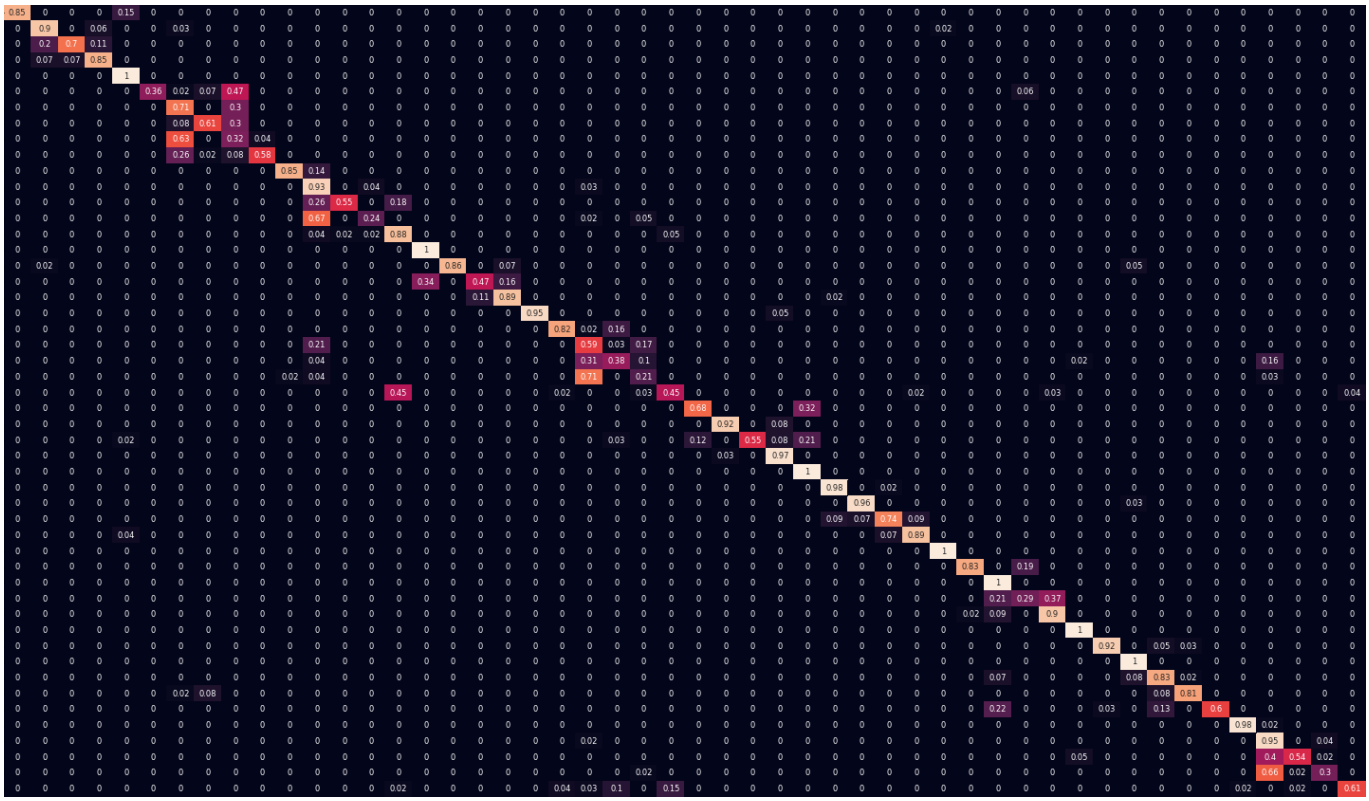


Figure A.1: Confusion Matrix produced by the pose-based classifier in single-label mode(Chapter 8, Table 8.2) for the NTU dataset

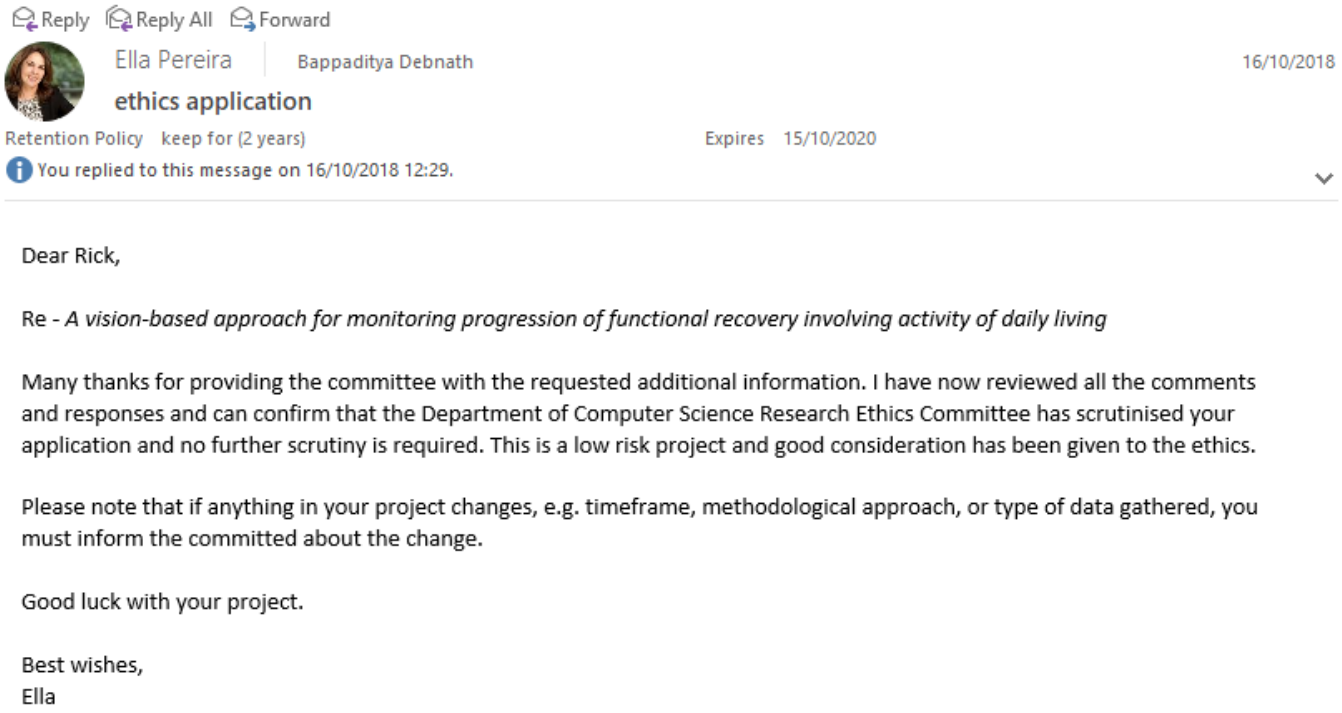
| | Split 1 | | | Split 2 | | Final |
|---|---|---|---|---|---|---|
| CSS | CST | Acc | CSS | CST | Acc | |
| 8 | 4 | 76.1 | 8 | 4 | 77.2 | 76.7 |
| 8 | 8 | 75.4 | 8 | 8 | **82.2** | **78.8** |
| 8 | 16 | 75.8 | 8 | 16 | 80.2 | 78.0 |
| 8 | 32 | 72.2 | 8 | 32 | 80.9 | 76.6 |
| 8 | 64 | 77.1 | 8 | 64 | 72.3 | 74.7 |
| 8 | 128 | **77.4** | 8 | 128 | 71.4 | 74.4 |
| 16 | 4 | 72.5 | 16 | 4 | 75.3 | 73.9 |
| 16 | 8 | 71.8 | 16 | 8 | 68.1 | 69.9 |
| 16 | 16 | 75.2 | 16 | 16 | 72.5 | 73.8 |
| 16 | 32 | 74.3 | 16 | 32 | 76.9 | 75.6 |
| 16 | 64 | 75.5 | 16 | 64 | 81.6 | 78.6 |
| 16 | 128 | 76.8 | 16 | 128 | 74.2 | 75.5 |
| 32 | 4 | 73.1 | 32 | 4 | 75.2 | 74.2 |
| 32 | 8 | 72.7 | 32 | 8 | 69.5 | 71.1 |
| 32 | 16 | 75.9 | 32 | 16 | 77.6 | 76.8 |
| 32 | 32 | 75.5 | 32 | 32 | 72.9 | 74.2 |
| 32 | 64 | 75.5 | 32 | 64 | 73.7 | 74.6 |
| 32 | 128 | 70.2 | 32 | 128 | 69.2 | 69.7 |
| 64 | 4 | 71.3 | 64 | 4 | 73.2 | 72.3 |
| 64 | 8 | 69.1 | 64 | 8 | 66.1 | 67.6 |
| 64 | 16 | 74.5 | 64 | 16 | 71.7 | 73.1 |
| 64 | 32 | 75.3 | 64 | 32 | 78.9 | 77.1 |
| 64 | 64 | 76.5 | 64 | 64 | 72.6 | 74.5 |
| 64 | 128 | 74.9 | 64 | 128 | 73.2 | 74.0 |
| 128 | 4 | 70.2 | 128 | 4 | 75.4 | 72.8 |
| 128 | 8 | 69.2 | 128 | 8 | 71.3 | 70.3 |
| 128 | 16 | 67.8 | 128 | 16 | 72.4 | 70.1 |
| 128 | 32 | 75.4 | 128 | 32 | 76.2 | 75.8 |
| 128 | 64 | 72.3 | 128 | 64 | 71.5 | 71.9 |
| 128 | 128 | 77.3 | 128 | 128 | 73.2 | 75.2 |

Table A.2: The table illustrates the impact of cluster size on the accuracy. CSS: Cluster Size Spatial Stream, CST: Cluster Size Temporal Stream, Acc. Accuracy

# Appendix B

# Ethical Clearance

## B.1 Approval



Reply Reply All Forward

Ella Pereira Bappaditya Debnath 16/10/2018

ethics application

Retention Policy keep for (2 years) Expires 15/10/2020

You replied to this message on 16/10/2018 12:29.

Dear Rick,

Re - *A vision-based approach for monitoring progression of functional recovery involving activity of daily living*

Many thanks for providing the committee with the requested additional information. I have now reviewed all the comments and responses and can confirm that the Department of Computer Science Research Ethics Committee has scrutinised your application and no further scrutiny is required. This is a low risk project and good consideration has been given to the ethics.

Please note that if anything in your project changes, e.g. timeframe, methodological approach, or type of data gathered, you must inform the committed about the change.

Good luck with your project.

Best wishes,
Ella

Figure B.1: Ethical clearance letter

## B.2 Participant Consent Form

**Project title:** A vision-based approach for monitoring progression of functional recovery involving activity of daily living

The project aims to identify and grade simple activities of daily living (ADL) from a video. Participants will be performing the following activities and will be filmed. The research team will guide the participants to perform the activities at 5 level of deftness. For e.g. slow and shaky hand movements while drinking water to steady and normal. The filmed content will be used as dataset to train an AI based model to identify the ADL and grade it. The dataset will be published and will be publicly available for other researchers to use once the project is completed.

List of ADL:

- Walking
- Drinking
- Standing up
- Sitting down
- Wearing glasses
- Reaching above
- Brushing floor
- Answering phone
- Clapping
- Brushing hair

**Researcher: Bappaditya Debnath (GTA)**

- I confirm that I have read and understand the information sheet for the above study
- I have been given the opportunity to ask questions
- I agree to take part in this study
- I understand that my participation is voluntary and that
- I am free to withdraw for up to four weeks after the date of the consent.

Name of participant ……
Signature ……
Date ……

Researcher ……
Signature ……
Date ……

## B.3 Participant Information Sheet

**Study title**

A vision-based approach for monitoring progression of functional recovery involving activity of daily living

**Principal researcher**

Bappaditya Debnath, PhD candidate, Department of Computer Science, Office: THF09, Phone: 01695 651872, Email: debnathb@edgehill.ac.uk

## B.4 Supervisors

:

Dr Ardhendu Behera, Senior Lecturer, Department of Computer Science, Office: THF14, Phone: 1695 657272, Email: beheraa@edgehill.ac.uk

Dr Swagat Kumar, Department of Computer Science, Office: PSS126, Phone: 01695 657417, Email: Kumars@edgehill.ac.uk

Prof Mary O'Brien, Department of Health and Social Care, Office: H217, Phone: 01695 650918, Email: obrienm@edgehill.ac.uk

**Invitation**

You are invited to take part in a Computer Science based research study. The study aims to research for a smart software that will enable automatic recognition of abnormal human movements while performing any normal activity (e.g: walking). Before deciding whether to take part, it is important for you to understand why the research is being done and what it will involve. You are advised to take time to read the information that follows carefully and discuss it with others if you wish. Please inform the researcher (Bappaditya Debnath) if you would like more information or if anything is unclear.

**What is the purpose of the study?**

The research is for my PhD. It aims to develop a software that can detect abnormalities in daily human movements from video footage of persons performing normal activities i.e. Activities of Daily Living (ADL). For e.g. a person recovering from stroke may not be able to walk properly. To research for such a software we need video samples of people walking or drinking water etc. which is not a normal and compare them with that of a normally performed ADL. The potential benefit for such research automated home monitoring for patients recovering from stroke, spinal cord injury etc.

**Why have I been invited?**

Video footage for such a research software needs to be in a studio environment where filming can be done with proper equipment. It is not possible to film actual patients performing such tasks in a studio environment. It is also not possible to asks patients to perform the following 10 activities, first in a normal manner and then like patients. Therefore, you are invited to perform the following 10 activities under the guidance of a qualified physiotherapist. First you are required to perform these activities normally as you would perform in you daily life. Then, you will be guided to act like a patient wherein you have to perform this activities like a patient would do. For e.g. while acting like putting you glasses on you will need to do it very slowly. List of ADL:

Walking, Drinking, Standing up, Sitting down, Wearing glasses, Reaching above, Brushing floor, Answering phone, Clapping, Brushing hair

**Consent**

You are requested to carefully go through this sheet of information (Participant Information Sheet), read all the information provided and ask for extra information if you need. You are not required to take part in the study and we seek your written consent through the participant consent for which should signed freely only after reading this information sheet and gathering all the information you want in addition to all the information we are giving you. By agreeing to take part in this research you agree to be filmed while performing the mentioned ADL activities. You also give consent to be filmed while performing the same ADLs while acting like a patient under supervision from a physiotherapist. There is no risk involved as you will be only doing everyday activities. The information being collected from you is a footage or video clip of you doing ADLs. The footage will be used to train and test the intelligent software algorithm that I need to develop for my PhD research. The clips will be only privately used by me during the entire duration of my PhD. After my research is published the videos will be made publicly available for the interest of larger research community. At Edge Hill, we are committed to respecting and protecting your personal information. To find ways in which we use your data, please see edgehill.ac.uk/about/legal/privacy. Data protection legislation & the lawful basis for processing personal data The University is committed to ensuring compliance with current data protection legislation and confirms that all data collected is used fairly, stored safely, and not disclosed to any other person unlawfully. The University is a data controller and, in some instances, may be a data processor of this data.

**Can I withdraw consent?**

You can withdraw your consent up to 30 days after recording. Beyond this period the data will be actively used for the research and upon completion of the research it will be shared publicly and therefore cannot be withdrawn. Because research is conducted in the public interest, participants will not have open-ended rights over their personal data under GDPR, although they retain the right to object.

**Will my participation be confidential?**

Except your video clip, that will show you performing ADLs, we are not collecting any data. The video clips will remain with me and will be kept encrypted at all times. But once the research is published the video clips will be publicly available for the interest of larger research community. Personal information data such as name, phone number etc. will not be collected. I am obliged to make a disclosure is made that suggests, either directly or indirectly, harm to the participant or to others, or criminal activity or bad practice. For this purpose clips containing the participants will be given an identification number. The number will be matched to the participant's name and shall be securely locked away in physical form.

**What will happen to the results of the research study?**

We intend to prepare a video Dataset, which is a collection of all the footages from all participants filmed in this study. The, dataset will be with me till the duration of my research and will not be shared. When the research is completed it will be published in a Computer Science related journals or conferences which will contain some sample images from the dataset. After publication the whole dataset will be publicly available for anyone to download in the interest of the larger research community. It is important for you to understand that by giving your consent and by agreeing to

take part in the research you also give your consent for your video clip to be publicly available after the completion of this research. The footage or the video clip however will be anonymous and NO other information such as your name, profession or any other personal details will be collected or published. Other than your video clip no other information is required for this research. The dataset description will also make it clear that the participants in the clips are actors performing ADLs and not actual patients. Once the research is published the Dataset will be publicly available for an indefinite period. Who has reviewed the study? The study has been reviewed by the relevant research ethics committee at Edge Hill University.

**Contact:**

Professor Ella Pereira, Department of Computer Science. Office: THG12, Phone: 01695 657639, Email: Pereirae@edgehill.ac.uk

**What will I be asked to do?**

You will be asked to come to our studio at computer science department wherein you will be performing the following ADL: List of ADL:

Walking
Drinking
Standing up
Sitting down
Wearing glasses
Reaching above
Brushing floor
Answering phone
Clapping
Brushing hair

You will be filmed 5 times while doing each of the above activities. 1 of these 5 will be normally what you would do in your daily life. For the other 4 times you will be acting like a person with movement difficulties. We aim to do it in one session and should take around 2 hours.

**What are the possible disadvantages and risks of taking part?**

Explain any possible side effects/adverse effects of taking part. This could include physical or psychological effects (if the subject matter is sensitive, embarrassing, or potentially upsetting). How will you respond should any adverse effects occur? What support systems will you have in place should it be a sensitive or upsetting topic, etc.? Are any safety measures on standby? How quickly will the participants be able to access the support (e.g. if you are referring people to a third party agency)?

**Health-related findings**

There is absolutely no health-related impact on participants or researchers for this research related activity. Edge Hill guidance on HRFs is available but, if you have any other questions, please consult this advice from the Medical Research Council or contact the Biological Safety Officer. What are

the possible benefits of taking part? You have the opportunity take part in a research that has the potential to improve home patient care. You will be paid @9 GBP per hour.

**Is there someone independent I can talk to about the research?**

You have the option to talk to Prof Ella Parreira, who is the Chairman of Computer Science ethics committee. Professor Ella Pereira, Department of Computer Science, Office: THG12, Phone: 01695 657639, Email: Pereirae@edgehill.ac.uk

**Support**

There is absolutely no risk involved to any of the participant or researchers as the activity requires you to perform simple activities of daily living while you are filmed.